Atrophy-centered subtyping of mild cognitive impairment

Kichang Kwak[1], Kelly S. Giovanello[1,2], Martin Styner[3,4] and Eran Dayan[1,5*] for the Alzheimer's Disease Neuroimaging Initiative[†]

1. Biomedical Research Imaging Center, University of North Carolina at Chapel Hill

2. Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill

3. Department of Computer Science, University of North Carolina at Chapel Hill

4. Department of Psychiatry, University of North Carolina at Chapel Hill

5. Department of Radiology, University of North Carolina at Chapel Hill

* ***Corresponding author***

Eran Dayan, Ph.D.

Address: 130 Mason Farm Road, CB 7513, Chapel Hill, NC, 27599

Tel: 919-843-8256

Email: eran_dayan@med.unc.edu

[†]Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of the ADNI and/or provided data but did not participate in analysis or writing of this article. 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

Mild cognitive impairment (MCI) is considered as the transitional phase between normal cognitive aging and Alzheimer's disease (AD). Nevertheless, trajectories of cognitive decline vary considerably among individuals with MCI. To address this heterogeneity, subtyping approaches have been developed, with the objective of identifying more homogenous subgroups and ultimately improving prognostic outcomes. To date, subtyping of MCI has been based primarily on cognitive performance measures, often resulting in indistinct boundaries between the proposed subgroups and limited validity. The degree to which markers of neurodegeneration such as brain atrophy can be used to subtype MCI into biologically and clinically meaningful subgroups remains unclear. Here we introduce and validate a data-driven subtyping method for MCI based solely upon measures of atrophy derived from structural magnetic resonance imaging (MRI). We trained a dense convolutional neural network to differentiate between patients with AD and age-matched cognitively normal (CN) subjects based on whole brain MRI features. We then deployed the trained model to classify individuals with MCI, as MCI-CN or MCI-AD, based on the degree to which their whole brain gray matter volume resembles CN-like or AD-like patterns. We subsequently validated the model-based subgroups using cognitive, clinical, fluid biomarker, and molecular neuroimaging data. Namely, we observed marked differences between the MCI-CN and MCI-AD groups in baseline and longitudinal cognitive and clinical rating scales, disease-free survival, cerebrospinal fluid (CSF) levels of amyloid beta and tau, fluorodeoxyglucose (FDG) and amyloid PET. Overall, the results suggest that patterns of atrophy in MCI are sufficiently distinct and heterogeneous, and can thus be used to subtype individuals into biologically and clinically meaningful subgroups.

Introduction

Mild cognitive impairment (MCI) is often construed as the transitional stage between normal aging-related cognitive decline and Alzheimer's disease (AD) *(1, 2)*. However, MCI is associated with marked etiological heterogeneity *(3)*. While the yearly risk of progression from MCI to AD is set at around 10% to 12% *(4–6)*, not all individuals with MCI eventually progress to AD and many demonstrate different outcomes, including the development of non-AD dementia or other neuropsychiatric conditions *(7, 8)*, or reversion to cognitively normal (CN) status *(9)*. Despite its ubiquity, the heterogeneity of MCI remains poorly understood, challenging further progress in research and care.

Attempts to constrain the heterogeneity of MCI via subtyping approaches have been proposed by several groups *(10–12)* almost exclusively relying on cognitive subtyping, that is, classification into subtypes which is based on subjects' performance in cognitive tests and tasks *(11, 13)*. For example, a common subtyping framework defines individuals with MCI as amnestic and non-amnestic, depending on whether or not memory loss is a predominant feature *(11)*. Other common subtyping approaches further classified MCI as being either single- or multiple-domain as a function of the number of cognitive domains where decline is observed *(14)*. More recently, studies have shown that a more comprehensive subtyping for MCI can be achieved based on similarities in neuropsychological test scores using clustering techniques *(15, 16)*.

While cognitive subtyping has been instrumental in delineating the various dimensions of cognitive performance that are affected in MCI, cognitive subtypes may suffer from insufficiently distinct boundaries and their validity has been questioned *(17)*. Moreover, a shift in the definition of preclinical, prodromal and clinical AD from a syndromic to a biological construct will hopefully facilitate advances in the identification of new treatment targets and to increased precision in interventional clinical trials *(18)*. Indeed, the recently proposed "AT(N)" framework for AD research *(19, 20)*, attempts to provides accurate, biologically-centered definitions for AD research based on multi-domain biomarkers for β-amyloid deposition ('A'), pathologic tau ('T'), and neurodegeneration ('N'). Nevertheless, biomarkers for neurodegeneration, particularly those based on structural magnetic resonance imaging (MRI), are not

specific to AD, and may be attributable to various other comorbidities *(18)*. Consequently, the degree to which heterogeneity in atrophy patterns in MCI can be leveraged to subtype individuals into homogeneous subgroups remain unclear.

In the current study, we tested whether patterns of brain atrophy, derived from MRI, are sufficient in allowing for the subtyping of MCI into biologically and clinically distinct subgroups. We propose and validate a novel data-driven subtyping approach based upon deep learning. We first train a dense convolutional neuronal network (CNN) to differentiate between AD and CN based on whole brain gray matter morphometric data. We then deploy the trained CNN to classify MCI subjects into two subgroups, MCI-AD and MCI-CN, corresponding to MCI subjects with AD-like and CN-like morphometric characteristics, respectively. We also identified the major regional atrophy patterns contributing to the differentiation between the two MCI subtypes through occlusion analysis *(21)*. The resulting labels were then validated against cerebrospinal fluid (CSF) biomarkers for β-amyloid (Aβ) and tau, baseline fluorodeoxyglucose (FDG) and amyloid positron emission tomography (PET), as well as baseline and longitudinal cognitive scores. Finally, we evaluated the degree of overlap between the modeling-based labels and those obtained through cognitive subtyping.

Results

Participant Characteristics

We analyzed data from 489 subjects, obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Our proposed modeling approach (see below) utilized AD and CN data for training, and MCI data for testing. Following the proposed NIA-AA guidelines *(18)*, AD subjects were included in the analysis if they displayed abnormal CSF $A\beta_{42}$ and p-tau_{181} levels (denoted henceforth as, A+T+). CN subjects were included in the analysis only if they displayed normal CSF $A\beta_{42}$ and p-tau_{181} levels (henceforth, A−T−). The demographic characteristics of subjects in the AD and CN groups are shown in Table 1. When comparing the AD, CN, and MCI groups (Table S1), there were significant difference in Clinical Dementia Rating (CDR: $F_{2,596} = 458.2$, $p < 0.001$), Alzheimer's Disease Assessment Scale (ADAS:

$F_{2,596}$ = 177.90, $p < 0.001$), and CSF biomarker concentrations (Aβ_{42}: $F_{2,596}$ = 146.45, $p < 0.001$, p-tau$_{181}$: $F_{2,596}$ = 86.08, $p < 0.001$). Post hoc comparisons with the Tukey test revealed significant pairwise difference (all p's < 0.001) between the 3 groups in all cognitive measures and CSF biomarker concentrations. No significant differences were observed between the groups in age ($F_{2,596}$ = 2.16, p = 0.12) and gender distribution (χ^2 = 3.06, p = 0.22).

A deep learning model for subtyping MCI subjects

Our major objective in the current study was to develop a deep learning modeling framework for subtyping MCI. To that end we utilized a dense CNN architecture (Fig. 1) *(22)*, which relies on whole brain gray matter (GM) density as input data. The model was first trained to differentiate AD and CN, with the assumption that these 2 distinct groups would provide the model with an adequate distribution of morphometric features sufficient for subtyping MCI subjects. Data augmentation was applied within the training data to increase its size and improve the performance of the model and its generalizability. We used 5-fold cross-validation within the training set to optimize and fine-tune the model's performance, finding similar performance across the different folds (Fig. S1). The model with the best performance achieved maximal accuracy of 93.75%, with an area under the curve (AUC) of the receiver operating characteristic (ROC) of 0.983 (Fig. S1). We subsequently deployed the model on the MCI data (Fig. 1), formalized as a binary classification problem with class labels AD and CN. Then, output values less than the default threshold of 0.5 were assigned to the class CN (henceforth, MCI-CN) and values greater than or equal to 0.5 were assigned to the class AD (henceforth, MCI-AD). We also validated the group differences across various thresholds for binary classification and observed relatively little differences as a function of threshold (Fig. S2).

-- Figure 1 Here --

Validation of model-based MCI subtypes with CSF biomarker concentrations and cognitive scores

Our data-driven approach for subtyping MCI resulted in two subgroups, MCI-AD and MCI-CN. We next assessed the validity of these two data-driven labels. We first compared the prevalence of the various biomarker profiles *(18)* in each of the subgroups, based on each subject's CSF $A\beta_{42}$ and p-tau_{181} biomarkers. Each subject was rated as either positive (i.e., abnormal) or negative (i.e., normal) in each biomarker, based on previously published cut-off values *(18)* (See Fig. S3). This resulted in 4 different profiles (A+T+, A+T−, A−T+, and A−T−, where 'A' denotes Aβ, and 'T' denotes tau) (Fig. 2A, B). The prevalence of the biomarker profiles differed significantly between the MCI-AD and MCI-CN subgroups (χ^2 = 21.40, p < 0.001). In particular, subjects with an abnormal CSF $A\beta_{42}$ or p-tau_{181} were more common in the MCI-AD (CSF $A\beta_{42}$: 71.6%; p-tau_{181}: 69.3%) than in the MCI-CN group (CSF $A\beta_{42}$: 50.6%; p-tau_{181}: 50.2%). Subjects with an abnormal CSF $A\beta_{42}$ and p-tau_{181} were more prominent in the MCI-AD (A+T+: 55.1%) than in the MCI-CN group (A+T+: 32.8%). In contrast, subjects with a normal CSF $A\beta_{42}$ and p-tau_{181} were more common in the MCI-CN (A−T−: 32.0%) than in MCI-AD group (A−T−: 14.2%).

We then investigated baseline differences in demographic characteristics, cognitive scores and continuous CSF concentrations between the MCI-AD and MCI-CN subgroups. There was a significant difference between the two subgroups in age (t_{378} = 6.44, p < 0.001), but not in gender distribution (χ^2 = 3.49, p = 0.06). Comparing cognitive scores between the MCI-AD and MCI-CN subgroups reveled significant differences in CDR (t_{378} = 3.46, p < 0.001; Fig 2C) and ADAS (t_{378} = 6.51, p < 0.001; See Fig S4) scores. All significant results were retained when controlling for age (all p values < 0.01). The comparison of CSF concentrations between the MCI-AD and MCI-CN subgroups revealed significant differences in CSF $A\beta_{42}$ (t_{378} = 4.55, p < 0.001; Fig 2D) and in p-tau_{181} (t_{378} = 3.81, p < 0.001; Fig 2E). The significant results were retained when controlling for age (both p values < 0.001).

-- Figure 2 Here --

Comparison of PET uptake in subtyped MCI

We next assessed group differences in PET uptake, focusing on Aβ- and FDG-PET. While Aβ-PET and CSF measures of Aβ are generally properly correlated with one another, a certain degree of discordance between the two markers has been consistently reported *(23, 24)*. Our analysis above revealed that the MCI-AD and MCI-CN subgroups differed in CSF $A\beta_{42}$ levels. We thus next aimed to complement this analysis by also comparing the subgroups across Aβ-PET. There were significant differences between the two subgroups in Aβ-PET (t_{377} = 4.36, $p < 0.001$; Fig 3A), with lower levels observed in the MCI-AD group, relative to the MCI-CN group. We have additionally examined group differences in FDG-PET uptake. Metabolic imaging studies utilizing FDG-PET for AD diagnosis are common *(25)*, and this modality has been proposed as a reliable and valid marker for neurodegeneration in AD *(18)*. We observed significant group differences in FDG-PET (t_{377} = 5.53, $p < 0.001$; Fig 3A), with lower uptake values obtained in the MCI-AD group. Group differences in both Aβ-PET and FDG-PET were retained after adjusting for the effect of age. We have additionally assessed group differences in PET uptake, comparing normal/abnormal uptake values after binarizing the data with established cut-off values *(26, 27)*. The prevalence of normal and abnormal Aβ-PET and FDG-PET differed between the two subtyped subgroups (Fig. S5). These differences were apparent in both Aβ-PET and FDG-PET (Aβ-PET: χ^2 = 12.35, $p < 0.001$; FDG-PET: χ^2 = 14.70, $p < 0.001$). In particular, there were more subjects with abnormal Aβ-PET in the MCI-AD (Aβ-PET+: 71.4%) than in MCI-CN group (Aβ-PET+: 52.6%). Similarly, abnormal FDG-PET was more common in the MCI-AD (FDG-PET+: 48.4%) than in MCI-CN group (FDG-PET+: 28.5%).

-- Figure 3 Here --

Contribution of brain regions in subtyped MCI: occlusion analysis

The results suggest that patterns of whole brain gray matter are sufficient for differentiating MCI subjects into 2 distinct subgroups. As our approach utilizes whole brain gray matter features, the major regional contributors to the model's output remain unclear. We thus next examined the relative lobar (frontal, parietal, medial temporal, lateral temporal, occipital, and cingulate) contribution to the performance of the

model, through occlusion analysis, as proposed previously *(21)*. Briefly, we retested the deep learning model iteratively, occluding a bilateral binary mask composed of each lobe from the model's test-set input data (Fig. 4A). This was achieved by setting the intensity values of each lobe to zero on each iteration. The Percentages of change in the model's output, with respect to the original results for classifying MCI subgroups were ranked and then compared across the different occluded lobes (See Fig. 4B). We found that the occlusion of the medial temporal and lateral temporal lobes led to dramatic changes in the model's output (Fig. 4C), relative to the original results. On the other hand, occlusion of the occipital and cingulate lobes had relatively little effect, resulting in model output that resembled the original results (Fig. 4C). Thus, the medial and lateral temporal lobes had the largest impact on the performance of the model and on the classification of MCI subjects into two distinct subgroups.

-- Figure 4 Here --

Longitudinal analysis of cognitive changes in the MCI subgroups

Our analysis reveals marked baseline differences between the MCI-AD and MCI-CN subgroups. These observed differences, nevertheless, cannot be taken to imply that the two MCI subgroups also differ in their prognostic outcomes. We next set out to evaluate if the MCI-AD and MCI-CN subgroups also exhibit differences in the progression to AD and in longitudinal cognitive performance. This analysis focused on subjects with data from at least 3 follow-up visits. Individual trajectories of longitudinal changes in cognitive performance varied between the two MCI subgroups (Fig. S6). Survival analysis *(28)* revealed marked differences between the two subgroups in their progression to AD (Log-rank test; $\chi^2 = 64.40$, $p < 0.001$), with the MCI-AD subgroup showing faster progression, relative to the MCI-CN subgroup, where slower progression was observed over time (Fig. 5A). We next used repeated measures-analysis of variance (RM-ANOVA) to compare changes in cognitive performance from baseline to the 2nd year follow-up visit, with group (MCI-AD, MCI-CN) as the between-subjects factor and time (baseline, follow-up) as the within-subject, repeated-measure factor. Focusing on CDR scores (Fig. 5B), this analysis revealed a significant interaction (group × time; $F_{1,358} = 14.92$, $p < 0.001$), along with significant main effects for group ($F_{1,358} =$

27.87, $p < 0.001$) and time ($F_{1,358} = 38.91$, $p < 0.001$). Thus, the MCI-AD group showed more pronounced changes in CDR scores between the testing sessions. Similarly, in the analysis of ADAS scores (See Fig. S7), a significant interaction (group × time; $F_{1,358} = 7.56$, $p < 0.001$) was observed, along with significant main effects for group ($F_{1,358} = 5.92$, $p = 0.02$), and time ($F_{1,358} = 87.12$, $p < 0.001$).

-- Figure 5 Here --

Concordance between the current MCI subtyping approach and cognitive subtyping

The data-driven approach proposed here subtypes MCI subjects solely based on patterns of whole brain gray matter volume. As the vast majority of existing subtyping approaches for MCI are based on cognitive profiles *(17)*, we wished to determine the extent of concordance between the current MCI subtyping approach and cognitive subtyping. We focused the comparison on a recently-introduced subtyping method where neuropsychological assessments are clustered into distinct subgroups *(15)*. We first clustered subjects (n=374 subjects who had available data) based on their neuropsychological assessments (Table S2). This resulted in 4 subgroups: *Dysnomic* (37.43% of subjects), *amnestic* MCI (36.63), *Dysexecutive* (6.95%), and *Cluster-Derived Normal* (18.98%) (Fig. 6A). These 4 subgroups showed significant group differences in the 6 neuropsychological assessments used for clustering ($p < 0.001$). Post-hoc comparisons with the Tukey multiple comparison test revealed that the *Dysnomic* group performed worse than all other groups in 5/6 measures of language. The *Dysexecutive* group performed worse than all other groups in assessments of attention/executive function, and the amnestic MCI group performed worse than the *Dysnomic* and *Cluster-Derived Normal* groups in assessments of memory function. Thus, the clustering of neuropsychological assessments resulted in distinct MCI subtypes, as previously reported *(15)*. We next compared the distributions of the neuropsychological subtypes within the MCI-AD and MCI-CN subgroups. We found significant differences between the two distributions ($\chi^2 = 30.45$, $p < 0.001$), observing more *Dysnomic* and *Dysexecutive* subjects in the MCI-AD (60%) than in the MCI-CN group (36.5%). On the other hand, as expected, the *Cluster-Derived Normal* profile, which shows the normal range of cores across all neuropsychological measures, was more common in the MCI-CN (24.9%) than in the MCI-AD group

(7.2%). Interestingly, the prevalence of the Amnestic MCI profile was similar between the MCI-AD (32.8%) and MCI-CN subgroups (38.6%). Altogether, while both the MCI-AD and MCI-CN subgroups displayed impairments in memory function, the former subgroup displayed more significant deficits in attention/executive and language function than the latter subgroup.

-- Figure 6 Here --

Discussion

Subtyping approaches for MCI have been proposed as a remedy for the large etiological heterogeneity characteristic of this elusive stage in cognitive aging. Current subtyping approaches primarily rely on neuropsychological profiles and may often result in blurred boundaries between subgroups and limited validity *(17)*. Here, we propose a novel data-driven subtyping approach, which utilizes CNNs to divide MCI subjects into subgroups based on the extent to which their brain atrophy patterns resemble those observed in AD as opposed to CN data. This approach resulted in two subgroups, MCI-AD and MCI-CN, denoting closer correspondence in gray matter patterns with the AD and CN, respectively. We then comprehensibly validated the model-based subgroups, findings marked group differences in baseline CSF biomarker concentrations and PET uptake, along with baseline and longitudinal cognitive performance scores. Through occlusion analysis *(21)* we investigated lobar contribution to the performance of the deep learning model, reporting that it mostly relied on gray matter volume from the medial and lateral temporal lobes. Finally, we found a limited degree of overlap between the current subtyping approach and that based on neuropsychological examination.

The purpose of the current study was to test whether distinct subgroups with differing structural brain atrophy patterns could be delineated within a heterogeneous clinical sample of individuals diagnosed with MCI. To that end, we utilized a deep learning framework, rather than other machine learning models, such as support vector machine, primarily since there has been a growing body of research demonstrating the

utility of deep learning models based on MRI-derived features in various tasks, such as diagnostic prediction *(29)*, image reconstruction *(30)* and segmentation *(31)*, and prognostic prediction of disease progression *(32)*. Our choice to utilize a deep learning framework was further motivated by the assumption that complex and non-linear relationships exist between whole brain structure and progression of MCI/AD. Similar to other machine and deep learning models where "transfer learning" *(33)* is applied, we propose a classification framework which is trained on a domain different than the one being tested *(34–36)*. However, rather than evaluating the performance of the model against clinically-defined labels (e.g., progressive and stable MCI, or AD convertors and non-convertors), our approach was to re-label data from individuals with MCI based on its proximity to the model's trained labels, that is, AD and CN.

We found robust differences between the MCI-AD and MCI-CN subgroups in CSF biomarker concentrations, cognitive scores, and PET uptake, suggesting our data-driven method subtypes MCI into biologically and clinically distinct subgroups. Moreover, the prevalence of biomarker profiles, defined based on established cut-off thresholds for the CSF $A\beta_{42}$ and p-tau_{181} biomarkers differed significantly between the MCI-AD and MCI-CN subgroups. Namely, abnormal CSF $A\beta_{42}$ (A+) and p-tau_{181} (T+) were more prevalent in the MCI-AD group. These findings are consistent with earlier studies where it was shown that positive CSF biomarker concentrations (AD-pathological) can predict conversion from MCI to AD with accuracy larger than 80% *(37, 38)*. We additionally found more pronounced cognitive impairment, as assessed with the CDR and ADAS scores, in the MCI-AD subgroup, relative to the MCI-CN subgroup. The two subtyped subgroups also exhibited marked differences in Aβ-PET, a marker of amyloid deposition, and FDG-PET, a commonly-used marker of neurodegeneration *(18)*. The topographical distribution of Aβ deposition, assessed with PET is predicative of progression of individuals with MCI to AD *(39)*, with abnormalities appearing long before the onset of clinical symptoms *(40)*. The Aβ-PET abnormalities observed in the MCI-AD subgroup, where longitudinal cognitive outcomes were poorer and progression to AD was quicker, are consistent with these findings. Similarly, the finding that negative FDG-PET was

more predominant in the MCI-CN subgroup is consistent with the observation that negative FDG-PET is highly predictive of clinically stable MCI *(41)*.

We complemented the initial analysis, which relied on whole brain gray matter patterns, with an analysis that combined deep learning classification with occlusion analysis *(21)*. This allowed us to identify the major lobar contributors to the performance of the subtyping model. The results revealed that the medial temporal, lateral temporal, and to a lesser extent the parietal lobe were more central to the model's performance than other lobes. These findings are consistent with earlier studies on AD pathology, where atrophy was reported in the hippocampus, amygdala, and entorhinal cortex *(42, 43)*. Our results also suggest that the occipital lobe played a more minor role in the performance of the model, consistent with Braak's staging scheme *(44)*, where the occipital lobe is shown to be affected only at later stages of AD *(45)*. Although our results highlight the central role of the medial and lateral temporal lobes in the subtyping of MCI, further examination into the possible involvement of other cortical and subcortical regions is warranted.

We examined the agreement between our atrophy-centered data-driven subtyping approach and that obtained based on neuropsychological assessments. The latter approach groups subjects based on the similarity of their neuropsychological assessments using clustering techniques, and thus goes beyond the more traditional *amnestic*/*non-amnestic* MCI subgrouping studied extensively in the literature *(46)*. However, neuropsychological and biomarker profiles are strongly heterogeneous in subjects who can be classified as having *amnestic* MCI *(47)*, and empirically-derived subtyping approaches depict heterogeneity that is not captured by conventional criteria. *(15)*. In our analysis, we found that the *Cluster-Derived normal* group, where cognitive function is unimpaired, was more predominant in the MCI-CN than the MCI-AD subgroups. (see Fig. 6). Moreover, while the *Amnestic* MCI profile showed a similar distribution in the two subgroups, *dysnomic* or *dysexecutive* subtypes were primarily represented in the MCI-AD subgroup. Overall, our findings demonstrate that the matching between the two subtyping approaches is incomplete.

Future research could attempt to combine the two approaches, to achieve neuropsychologically distinct subgroups, that show differing patterns of brain atrophy.

Several limitations should be noted when considering the current results. First, in this proof-of-concept stage we used atrophy patterns to subtype MCI into two distinct subgroups. We acknowledge that a larger number of subgroups would be needed to better capture the heterogeneity of MCI. In principle, our approach could be modified to output more than two subgroups, however, we believe that a larger number of MCI subjects than that used here would be needed in order to achieve robust and generalizable results. Second, as also noted above, we did not attempt to combine neuropsychological and atrophy-centered features in the current study, primarily as our major motivation was to examine if the latter type of features is sufficient in subtyping MCI. Future research could extend our approach to test if a combination of cognitive and neurobiological features better captures the heterogeneity of MCI.

In conclusion, the current study demonstrates that patterns of gray matter atrophy are sufficient for subtyping MCI into biologically and clinically distinct subgroups. These results further highlight the need to consider the heterogeneity of MCI when attempting to understand the pathological mechanisms of dementia, while providing a potential tool for individualized disease prognosis.

Materials and Methods

Subjects

Data used in the preparation of this study were obtained from the ADNI. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see http://www.adni-info.org/. We used AD (n=110) and CN (n=109) data to train the proposed deep learning model. Then, we subtyped MCI (n=380) into subgroups, validating the model's output with cognitive scores, CSF biomarker levels and PET uptake available in the same subjects. One subject had no

FDG-PET uptake data, while 20 subjects were excluded from the longitudinal analysis of cognitive performance since they had data in less than 3 testing time points. Missing data at the 2nd year follow-up was imputed using linear interpolation/extrapolation. All subjects provided written informed consent and the study's protocol was approved by the local Institutional Review Boards.

Study design

In this study, we subtyped MCI subjects using a deep learning approach, based on patterns of brain structural atrophy, derived from MRI. We then validated the model's output (i.e., subgroups) using CSF biomarker, PET and cognitive/clinical data. Specifically, we trained a dense CNN *(48),* to differentiate data from AD and CN subjects based on whole brain atrophy patterns. To provide the model with adequate well-defined training data, all AD subjects where A+ and T+, that is, they had abnormal levels of CSF $A\beta_{42}$ and p-tau181. CN subjects were all A- and T-, in other words, they showed normal levels in the same CSF biomarkers. Previously determined cutoff values *(18)* for abnormal Aβ42 (Aβ42 < 976.6pg/ml) and p-tau181 (p-tau181 > 21.8pg/ml) were used. We reasoned that training the model with well-differentiated AD (A+ T+) and CN (A- T-) data would allow the model to learn a more discriminative set of pathological features. We then deployed the trained CNN to classify MCI subjects (n=380), as either AD-like or CN-like. We subsequently validated the model's output labels with baseline molecular and metabolic neuroimaging, cognitive scales and tests and CSF biomarkers. We additionally examined longitudinal changes in cognitive scores (between baseline and the 2nd year follow-up visit, which was the latest visit where imputation for missing data could be achieved reliably) and calculated disease-free survival in the MCI-AD and MCI-CN subgroups, to assess differences in the progression to AD. Finally, we identified the major regions (lobes) contributing to the performance of the model through occlusion analysis *(21)* (see below), and assessed the intersection between the modeling-based labels and those obtained through cognitive subtyping.

Image acquisition

Structural MRI (T1) data used in the deep learning model were acquired at ADNI sites using 3T scanners and were based on either an inversion recovery-fast spoiled gradient recalled (IR-SPGR) or magnetization-prepared rapid gradient-echo (MP-RAGE) sequences *(49)*. Aβ-PET data were acquired in a 20-min dynamic emission scan, composed of four 5-min frames. The data was acquired 50–70 min after the injection of 10.0 mCi of [^{18}F]-AV45. FDG-PET data were acquired in a 30-min dynamic emission scan, composed of six 5-min frames. The data was acquired 30–60 min after the injection of 5.0 mCi of [^{18}F]-FDG. PET data ran through a strict quality control procedure to assess image quality. Standardized image preprocessing correction steps were applied to produce uniform data across the ADNI PET cores. These steps included frame co-registration, averaging across the dynamic range, and standardization with respect to orientation, voxel size, and intensity. Full details of the T1 and PET acquisition parameters and imaging processing steps are listed on the ADNI website (http://adni.loni.usc.edu/methods/).

CSF collection

CSF collection, shipping, aliquoting, storage as well as analysis followed ADNI's standardized procedures (http://www.adni-info.org/) *(50, 51)*. Collected CSF samples were frozen on dry ice right after collection (within 1 hour). The samples were then shipped overnight, also on dry ice, to the ADNI Biomarker Core laboratory at the University of Pennsylvania. Aliquots (0.5mL) were prepared from the CSF samples and were then stored at −80℃ in barcode-labeled polypropylene vials. Samples for $A\beta_{42}$ and p-tau_{181} were then measured using Elecsys immunoassays. The lower and upper technical limits for the Elecsys $A\beta_{42}$ CSF immunoassay were 200 to 1700pg/mL. The limits for the Elecsys p-tau_{181} CSF immunoassay were 8 to 120pg/mL

Image processing

The deep learning model used for subtyping MCI data utilized whole brain gray matter data. MRI data were analyzed using Statistical Parametric Mapping 12 (SPM12; Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm) running on MATLAB 9.8.0 (Math-Works, Natick, MA, USA). Briefly, all MR images were aligned to the anterior

commissure and segmented into gray matter (GM), white matter and CSF using the unified segmentation procedure *(52)*, implemented in SPM12. To improve the registration of the GM maps, we used the diffeomorphic anatomic registration through an exponentiated lie algebra algorithm (DARTEL) *(53)*. This resulted in more precise spatial normalization to the template. The DARTEL used subject-specific deformation fields to warp the GM map into subject-specific space, resampled at 2mm isotropic voxels. Then the warped GM maps were affine transformed into Montreal Neurological Institute (MNI) space. In addition to using whole brain GM volume data, we evaluated lobar contribution to the performance of the deep learning model through occlusion analysis (see below).

PET imaging data was used to validate the MCI subtypes outputted by the deep learning model. Detailed acquisition and standardized pre-processing procedures used with the PET images are available at the ADNI website (http://adni.loni.usc.edu/methods/). Amyloid PET uptake was calculated by averaging across 4 cortical regions (frontral, anterior cingulate, precuneus, and parietal cortex) relative to the whole cerebellum region *(26)*. Similarly, FDG-PET uptake was calculated by averaging across a set of pre-defined regions (angular gyrus, posterior cingulate, inferior temporal gyrus) relative to pons/vermis reference regions *(57)*.

Deep learning model architecture

Model architecture: The deep learning model used for subtyping of MCI (See Fig. 1) was based on the DenseNet architecture *(22)*. It consisted of a convolutional layer, 4 dense blocks, 3 transition layers, a global averaging pooling layer and a fully-connected layer. This state-of-the-art convolutional neural network architecture was chosen as it shows excellent feature propagation and classification performance while alleviating the vanishing gradient problem and significantly reducing the number of parameters used by the model *(22)*. First, the whole brain image with dimensions of 91×109×91 was passed through a stack of convolutional layers, where the filters were of size 5×5×5. The convolution stride was set to 1 voxel, while the size of the max-pooling layer was 2×2×2, with a kernel size set at 2×2×2. The dense block consisted of multiple convolution units, which were equipped with a batch normalization layer, leaky rectified linear unit, a 1×1×1 convolutional layer, a 5×5×5 convolutional layer and a dropout layer. Every convolutional

unit was connected to all the previous layers via shortcut connections. Dimensionality reduction of feature maps between dense blocks was achieved through the transition layer. The transition layer included a batch normalization layer, a leaky rectified linear unit, a convolutional layer of size 1×1×1, and an averaging pooling layer of size 2×2×2. The global averaging pooling layer was then concatenated and connected through a fully-connected layer. The model's output values were processed by the fully-connected layer which used a sigmoid activation function. The output layer mapped all values greater than 0.5 as 1 (positive class: MCI-AD) and all values less than or equal to 0.5 as 0 (negative class: MCI-CN), while testing MCI data based on the pre-trained model (Fig. 1).

Implementation

The Keras application programming interface in TensorFlow 2.0 was used for building the deep learning model. Model training and testing were performed in a parallelized manner with an Ubuntu 18.04.3 operating system, utilizing two Nvidia Tesla V100 graphic cards with 16GB memory each. The model was trained with a mini-batch size of 24 and 200 epochs, and optimized using stochastic gradient descent based on adaptive estimation of first- and second-order moments *(58)* and an exponentially decaying learning rate. The initial learning rate was set at 0.0001 and decayed by 0.9 after every 10000 steps. A dropout layer was added to the dense block, with the dropout rate set to 0.2. In the batch normalization step, beta and gamma weights were initialized with L2 regularization set at 1×10^{-4} and epsilon set to 1.1×10^{-5}. In the fully-connected layer, the L2 regularization penalty coefficient was set at 0.01. We observed stability in the model after an iteration of 150 epochs.

Occlusion analysis

To identify the more specific regional contribution to the performance of the deep learning model and the differentiation between the two MCI subgroups, we integrated occlusion analysis (e.g., *(59)*) into the classification framework. Cortical regions were first segmented with an automated segmentation tool available in FreeSurfer v6.0 (https://surfer.nmr.mgh.harvard.edu/), resulting in a parcellation of the cerebral cortex into 34 sulcul and gyral regions of interest (ROI) per hemisphere, according to the Desikan-Killiany

protocol *(54, 55)*. We then merged individual ROIs to 6 lobes: frontal, parietal, medial temporal, lateral temporal, occipital, and cingulate *(56)*. These ROIs were then masked out (setting their voxels to zero) from the testing phase's input data (Fig. 4A). We evaluated the performance of the different models (i.e., with each occluded lobe), relative to the intact model, quantifying the percentage of change in the model's accuracy with reference to the original results.

Clustering based on neuropsychological assessments

Neuropsychological testing was comprised of six measures assessing three different cognitive domains: (1) Memory: Rey auditory verbal learning test (RAVLT) *(60)*, 30-minute delayed free recall and the RAVLT recognition test. (2) Language: Animal fluency *(55)* and the 30-item Boston naming test total score *(62)*. (3) Attention/Executive: Trail making test (TMT) *(63)*, part A and part B. Similar to previous research *(15)* raw scores in these neuropsychological measures were used to cluster MCI subjects into subgroups. Raw scores were first transformed into age- and education-adjusted z-scores based on means and standard deviations in each measure, calculated in the CN group. Then, an agglomerative hierarchical clustering analysis was performed on the z-scores using Ward's method *(64)*. The clustering analysis resulted in four distinct subgroups *(15)*: *amnestic* MCI, *dysnomic* MCI, *dysexecutive* MCI, and *cluster-derived normal* group.

Statistical analysis and visualization

All analyses were performed using the R statistical software, version 4.0.2 (https://www.r-project.org). Group differences in continuous variables were analyzed with an ANOVA or with independent-sample t-tests. ANOVAs were followed, when relevant, by Tukey post-hoc comparisons. Chi-squared tests were applied to evaluate differences in categorical variables. Longitudinal data were analyzed using the R WRS2 package *(65)*, and were based on a RM-ANOVA, with group (MCI-AD, MCI-CN) serving as the between-subjects factor and time (baseline, follow-up) as the within-subject factor. Mauchly's test was applied to test for violations in the assumption of sphericity, followed by Greenhouse-Geisser corrections, if necessary. The plots for comparing cognitive scores, PET, and CSF between subgroups were based on the R

RainCloudPlots package *(66)*. Regional imaging results were displayed on a surface using the R ggseg (https://lcbc-uio.github.io/ggseg/) package.

References

1. R. C. Petersen, R. O. Roberts, D. S. Knopman, B. F. Boeve, Y. E. Geda, R. J. Ivnik, G. E. Smith, C. R. Jack, Mild cognitive impairment: Ten years later, *Arch. Neurol.* **66**, 1447–1455 (2009).

2. J. J. Manly, M. X. Tang, N. Schupf, Y. Stern, J. P. G. Vonsattel, R. Mayeux, Frequency and course of mild cognitive impairment in a multiethnic community, *Ann. Neurol.* **63**, 494–506 (2008).

3. C. DeCarli, Mild cognitive impairment: Prevalence, prognosis, aetiology, and treatment, *Lancet Neurol.* **2**, 15–21 (2003).

4. J. Bowen, L. Teri, W. Kukull, W. McCormick, S. M. McCurry, E. B. Larson, Progression to dementia in patients with isolated memory loss, *Lancet* **349**, 763–765 (1997).

5. R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, E. Kokmen, Mild cognitive impairment: Clinical characterization and outcome, *Arch. Neurol.* **56**, 303–308 (1999).

6. R. C. Petersen, J. C. Stevens, M. Ganguli, E. G. Tangalos, J. L. Cummings, S. T. DeKosky, Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review), *Neurology* **56**, 1133–1142 (2001).

7. J. S. Meyer, G. Xu, J. Thornby, M. H. Chowdhury, M. Quach, Is mild cognitive impairment prodromal for vascular dementia like Alzheimer's disease?, *Stroke* **33**, 1981–1985 (2002).

8. T. J. Ferman, G. E. Smith, K. Kantarci, B. F. Boeve, V. S. Pankratz, D. W. Dickson, N. R. Graff-Radford, Z. Wszolek, J. Van Gerpen, R. Uitti, O. Pedraza, M. E. Murray, J. Aakre, J. Parisi, D. S. Knopman, R. C. Petersen, Nonamnestic mild cognitive impairment progresses to dementia with Lewy bodies, *Neurology* **81**, 2032–2038 (2013).

9. T. D. Koepsell, S. E. Monsell, Reversion from mild cognitive impairment to normal or near-Normal cognition; Risk factors and prognosis, *Neurology* **79**, 1591–1598 (2012).

10. M. Grundman, R. C. Petersen, S. H. Ferris, R. G. Thomas, P. S. Aisen, D. A. Bennett, N. L. Foster, C. R. Jack, D. R. Galasko, R. Doody, J. Kaye, M. Sano, R. Mohs, S. Gauthier, H. T. Kim, S. Jin, A. N. Schultz, K. Schafer, R. Mulnard, C. H. Van Dyck, J. Mintzer, E. Y. Zamrini, D. Cahn-Weiner, L. J. Thal, Mild Cognitive Impairment Can Be Distinguished from Alzheimer Disease and Normal Aging for Clinical Trials, *Arch. Neurol.* **61**, 59–66 (2004).

11. B. Winblad, K. Palmer, M. Kivipelto, V. Jelic, L. Fratiglioni, L. O. Wahlund, A. Nordberg, L. Bäckman, M. Albert, O. Almkvist, H. Arai, H. Basun, K. Blennow, M. De Leon, C. Decarli, T. Erkinjuntti, E. Giacobini, C. Graff, J. Hardy, C. Jack, A. Jorm, K. Ritchie, C. Van Duijn, P. Visser, R. C. Petersen, Mild cognitive impairment - Beyond controversies, towards a consensus: Report of the International Working Group on Mild Cognitive Impairment, *J. Intern. Med.* **256**, 240–246 (2004).

12. M. H. Tabert, J. J. Manly, X. Liu, G. H. Pelton, S. Rosenblum, M. Jacobs, D. Zamora, M. Goodkind, K. Bell, Y. Stern, D. P. Devanand, Neuropsychological prediction of conversion to alzheimer disease in patients with mild cognitive impairment, *Arch. Gen. Psychiatry* **63**, 916–924 (2006).

13. B. Reisberg, S. H. Ferris, M. J. de Leon, E. S. E. Franssen, A. Kluger, P. Mir, J. Borenstein, A. E. George, E. Shulman, G. Steinberg, J. Cohen, Stage-specific behavioral, cognitive, and in vivo changes in community residing subjects with age-associated memory impairment and primary degenerative dementia of the Alzheimer type, *Drug Dev. Res.* **15**, 101–114 (1988).

14. R. C. Petersen, R. Doody, A. Kurz, R. C. Mohs, J. C. Morris, P. V. Rabins, K. Ritchie, M. Rossor, L. Thal, B. Winblad, Current concepts in mild cognitive impairment, *Arch. Neurol.* **58**, 1985–1992 (2001).

15. E. C. Edmonds, L. Delano-Wood, L. R. Clark, A. J. Jak, D. A. Nation, C. R. McDonald, D. J. Libon, R. Au, D. Galasko, D. P. Salmon, M. W. Bondi, Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors, *Alzheimers Dement* **11**, 415–424 (2015).

16. M. M. Machulda, E. S. Lundt, S. M. Albertson, W. K. Kremers, M. M. Mielke, D. S. Knopman, M. W. Bondi, R. C. Petersen, Neuropsychological subtypes of incident mild cognitive impairment in the Mayo Clinic Study of Aging, *Alzheimers Dement* **15**, 878–887 (2019).

17. T. F. Hughes, B. E. Snitz, M. Ganguli, Should mild cognitive impairment be subtyped?, *Curr. Opin. Psychiatry* **24**, 237–242 (2011).

18. C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J. L. Molinuevo, T. Montine, C. Phelps, K. P. Rankin, C. C. Rowe, P. Scheltens, E. Siemers, H. M. Snyder, R. Sperling, C. Elliott, E. Masliah, L. Ryan, N. Silverberg, NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease, *Alzheimers Dement* **14**, 535–562 (2018).

19. B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. Dekosky, S. Gauthier, D. Selkoe, R. Bateman, S. Cappa, S. Crutch, S. Engelborghs, G. B. Frisoni, N. C. Fox, D. Galasko, M. O. Habert, G. A. Jicha, A. Nordberg, F. Pasquier, G. Rabinovici, P. Robert, C. Rowe, S. Salloway, M. Sarazin, S. Epelbaum, L. C. de Souza, B. Vellas, P. J. Visser, L. Schneider, Y. Stern, P. Scheltens, J. L. Cummings, Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria, *Lancet Neurol.* **13**, 614–629 (2014).

20. M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies, C. H. Phelps, 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, *Focus (Madison).* **11**, 96–106 (2013).

21. K. Kwak, M. Niethammer, K. S. Giovanello, M. Styner, E. Dayan, The contribution of hippocampal subfields to the progression of neurodegeneration., *bioRxiv* (2020).

22. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, (2017).

23. S. M. Landau, M. Lu, A. D. Joshi, M. Pontecorvo, M. A. Mintun, J. Q. Trojanowski, L. M. Shaw, W. J. Jagust, for the A. D. N. Initiative, Comparing PET imaging and CSF measurements of Aβ, *Ann. Neurol.* **74**, 826–836 (2013).

24. A. De Wilde, J. Reimand, C. E. Teunissen, M. Zwan, A. D. Windhorst, R. Boellaard, W. M. Van Der Flier, P. Scheltens, B. N. M. Van Berckel, F. Bouwman, R. Ossenkoppele, Discordant amyloid-β PET and CSF biomarkers and its clinical consequences, *Alzheimer's Res. Ther.* **11**, 78 (2019).

25. L. Mosconi, P. F. McHugh, FDG- and amyloid-PET in Alzheimer's disease: Is the whole greater than the sum of the parts?, *Q. J. Nucl. Med. Mol. Imaging* **55**, 250–264 (2011).

26. S. M. Landau, M. A. Mintun, A. D. Joshi, R. A. Koeppe, R. C. Petersen, P. S. Aisen, M. W. Weiner, W. J. Jagust, Amyloid deposition, hypometabolism, and longitudinal cognitive decline, *Ann. Neurol.* **72**, 578–586 (2012).

27. S. M. Landau, D. Harvey, C. M. Madison, E. M. Reiman, N. L. Foster, P. S. Aisen, R. C. Petersen, L. M. Shaw, J. Q. Trojanowski, C. R. Jack, M. W. Weiner, W. J. Jagust, Comparing predictors of conversion and decline in mild cognitive impairment, *Neurology* **75**, 230–238 (2010).

28. E. L. Kaplan, P. Meier, Nonparametric Estimation from Incomplete Observations, *J. Am. Stat. Assoc.* **53**, 457–481 (1958).

29. J. Wang, M. J. Knol, A. Tiulpin, F. Dubost, M. De Bruijne, M. W. Vernooij, H. H. H. Adams, M. A. Ikram, W. J. Niessen, G. V. Roshchupkin, Gray matter age prediction as a biomarker for risk of dementia, *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21213–21218 (2019).

30. G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, D. Firmin, DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction, *IEEE Trans. Med. Imaging* **37**, 1310–1321 (2018).

31. N. Nogovitsyn, R. Souza, M. Muller, A. Srajer, S. Hassel, S. R. Arnott, A. D. Davis, G. B. Hall, J. K. Harris, M. Zamyadi, P. D. Metzak, Z. Ismail, S. L. Bray, C. Lebel, J. M. Addington, R. Milev, K. L. Harkness, B. N. Frey, R. W. Lam, S. C. Strother, B. I. Goldstein, S. Rotzinger, S. H. Kennedy, G. M. MacQueen, Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants, *Neuroimage* **197**, 589–597 (2019).

32. E. Hosseini-Asl, R. Keynton, A. El-Baz, in *Proceedings - International Conference on Image Processing, ICIP*, (2016).

33. H. Greenspan, B. Van Ginneken, R. M. Summers, Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique, *IEEE Trans. Med. Imaging* **35**, 1153–1159 (2016).

34. B. Cheng, M. Liu, D. Zhang, B. C. Munsell, D. Shen, Domain Transfer Learning for MCI Conversion Prediction, *IEEE Trans. Biomed. Eng.* **62**, 1805–1817 (2015).

35. H. Li, M. Habes, D. A. Wolk, Y. Fan, A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data, *Alzheimers Dement* **15**, 1059–1070 (2019).

36. C. Y. Wee, C. Liu, A. Lee, J. S. Poh, H. Ji, A. Qiu, Cortical graph neural network for AD and MCI diagnosis and transfer learning across populations, *NeuroImage Clin.* **23**, 101929 (2019).

37. O. Hansson, H. Zetterberg, P. Buchhave, E. Londos, K. Blennow, L. Minthon, Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: A follow-up study, *Lancet Neurol.* **5**, 228–234 (2006).

38. N. Mattsson, H. Zetterberg, O. Hansson, N. Andreasen, L. Parnetti, M. Jonsson, S. K. Herukka, W. M. Van Der Flier, M. A. Blankenstein, M. Ewers, K. Rich, E. Kaiser, M. Verbeek, M. Tsolaki, E. Mulugeta, E. Rosén, D. Aarsland, P. Jelle Visser, J. Schröder, J. Marcusson, M. De Leon, H. Hampel, P. Scheltens, T. Pirttilä, A. Wallin, M. Eriksdotter Jönhagen, L. Minthon, B. Winblad, K. Blennow, CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment, *JAMA - J. Am. Med. Assoc.* **302**, 385–393 (2009).

39. T. A. Pascoal, J. Therriault, S. Mathotaarachchi, M. S. Kang, M. Shin, A. L. Benedet, M. Chamoun, C. Tissot, F. Lussier, S. Mohaddes, J. Soucy, G. Massarweh, S. Gauthier, P. Rosa-Neto, Topographical distribution of Aβ predicts progression to dementia in Aβ positive mild cognitive impairment, *Alzheimer's Dement. Diagnosis, Assess.*

Dis. Monit. **12**, e12037 (2020).

40. W. J. Jansen, R. Ossenkoppele, D. L. Knol, B. M. Tijms, P. Scheltens, F. R. J. Verhey, P. J. Visser, P. Aalten, D. Aarsland, D. Alcolea, M. Alexander, I. S. Almdahl, S. E. Arnold, I. Baldeiras, H. Barthel, B. N. M. Van Berckel, K. Bibeau, K. Blennow, D. J. Brooks, M. A. Van Buchem, V. Camus, E. Cavedo, K. Chen, G. Chetelat, A. D. Cohen, A. Drzezga, S. Engelborghs, A. M. Fagan, T. Fladby, A. S. Fleisher, W. M. Van Der Flier, L. Ford, S. Forster, J. Fortea, N. Foskett, K. S. Frederiksen, Y. Freund-Levi, G. B. Frisoni, L. Froelich, T. Gabryelewicz, K. D. Gill, O. Gkatzima, E. Gomez-Tortosa, M. F. Gordon, T. Grimmer, H. Hampel, L. Hausner, S. Hellwig, S. K. Herukka, H. Hildebrandt, L. Ishihara, A. Ivanoiu, W. J. Jagust, P. Johannsen, R. Kandimalla, E. Kapaki, A. Klimkowicz-Mrowiec, W. E. Klunk, S. Kohler, N. Koglin, J. Kornhuber, M. G. Kramberger, K. Van Laere, S. M. Landau, D. Y. Lee, M. De Leon, V. Lisetti, A. Lleo, K. Madsen, W. Maier, J. Marcusson, N. Mattsson, A. De Mendonca, O. Meulenbroek, P. T. Meyer, M. A. Mintun, V. Mok, J. L. Molinuevo, H. M. Mollergard, J. C. Morris, B. Mroczko, S. Van Der Mussele, D. L. Na, A. Newberg, A. Nordberg, A. Nordlund, G. P. Novak, G. P. Paraskevas, L. Parnetti, G. Perera, O. Peters, J. Popp, S. Prabhakar, G. D. Rabinovici, I. H. G. B. Ramakers, L. Rami, C. R. De Oliveira, J. O. Rinne, K. M. Rodrigue, E. Rodriguez-Rodriguez, C. M. Roe, U. Rot, C. C. Rowe, E. Ruther, O. Sabri, P. Sanchez-Juan, I. Santana, M. Sarazin, J. Schroder, C. Schutte, S. W. Seo, F. Soetewey, H. Soininen, L. Spiru, H. Struyfs, C. E. Teunissen, M. Tsolaki, R. Vandenberghe, M. M. Verbeek, V. L. Villemagne, S. J. B. Vos, L. J. C. Van Waalwijk Van Doorn, G. Waldemar, A. Wallin, A. K. Wallin, J. Wiltfang, D. A. Wolk, M. Zboch, H. Zetterberg, Prevalence of cerebral amyloid pathology in persons without dementia: A meta-analysis, *JAMA* **313**, 1924–1938 (2015).

41. L. Iaccarino, A. Sala, D. Perani, Predicting long-term clinical stability in amyloid-positive subjects by FDG-PET, *Ann. Clin. Transl. Neurol.* **6**, 1113–1120 (2019).

42. D. J. A. Callen, S. E. Black, F. Gao, C. B. Caldwell, J. P. Szalai, Beyond the hippocampus: MRI volumetry confirms widespread limbic atrophy in AD, *Neurology* **57**, 1669–1674 (2001).

43. C. R. Jack, M. M. Shiung, J. L. Gunter, P. C. O'Brien, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. J. Ivnik, G. E. Smith, R. H. Cha, E. G. Tangalos, R. C. Petersen, Comparison of different MRI brain athrophy rate measures with clinical disease progression in AD, *Neurology* **62**, 591–600 (2004).

44. H. Braak, E. Braak, Neuropathological stageing of Alzheimer-related changes, *Acta Neuropathol.* **82**, 239–259 (1991).

45. H. Braak, I. Alafuzoff, T. Arzberger, H. Kretzschmar, K. Tredici, Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry, *Acta Neuropathol.* **112**, 389–404 (2006).

46. L. R. Clark, L. Delano-Wood, D. J. Libon, C. R. Mcdonald, D. A. Nation, K. J. Bangen, A. J. Jak, R. Au, D. P. Salmon, M. W. Bondi, Are empirically-derived subtypes of mild cognitive impairment consistent with conventional subtypes?, *J. Int. Neuropsychol. Soc.* **19**, 635–645 (2013).

47. J. Nettiksimmons, C. DeCarli, S. Landau, L. Beckett, Biological heterogeneity in ADNI amnestic mild cognitive impairment, *Alzheimers Dement* **10**, 511–521 (2014).

48. G. Huang, Z. Liu, L. Van der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, *CVPR* (2017).

49. C. R. Jack, M. A. Bernstein, B. J. Borowski, J. L. Gunter, N. C. Fox, P. M. Thompson, N. Schuff, G. Krueger, R. J. Killiany, C. S. Decarli, A. M. Dale, O. W. Carmichael, D. Tosun, M. W. Weiner, Update on the Magnetic Resonance Imaging core of the Alzheimer's Disease Neuroimaging Initiative, *Alzheimers Dement* **6**, 212–220 (2010).

50. L. M. Shaw, H. Vanderstichele, M. Knapik-Czajka, C. M. Clark, P. S. Aisen, R. C. Petersen, K. Blennow, H. Soares, A. Simon, P. Lewczuk, R. Dean, E. Siemers, W. Potter, V. M. Y. Lee, J. Q. Trojanowski, Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects, *Ann. Neurol.* **65**, 403–413 (2009).

51. J. Q. Trojanowski, H. Vandeerstichele, M. Korecka, C. M. Clark, P. S. Aisen, R. C. Petersen, K. Blennow, H. Soares, A. Simon, P. Lewczuk, R. Dean, E. Siemers, W. Z. Potter, M. W. Weiner, C. R. Jack, W. Jagust, A. W. Toga, V. M. Y. Lee, L. M. Shaw, Update on the biomarker core of the Alzheimer's Disease Neuroimaging Initiative subjects, *Alzheimers Dement* **6**, 230–238 (2010).

52. J. Ashburner, K. J. Friston, Unified segmentation, *Neuroimage* **26**, 839–851 (2005).

53. J. Ashburner, A fast diffeomorphic image registration algorithm, *Neuroimage* **38**, 95–113 (2007).

54. B. Fischl, A. Van Der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. H. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, V. Caviness, N. Makris, B. Rosen, A. M. Dale, Automatically Parcellating the Human Cerebral Cortex, *Cereb. Cortex* **14**, 11–22 (2004).

55. R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, R. J. Killiany, An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, *Neuroimage* **31**, 968–980 (2006).

56. A. Klein, J. Tourville, 101 labeled brain images and a consistent human cortical labeling protocol, *Front. Neurosci.* **6** (2012).

57. S. M. Landau, D. Harvey, C. M. Madison, R. A. Koeppe, E. M. Reiman, N. L. Foster, M. W. Weiner, W. J. Jagust, Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI, *Neurobiol. Aging* **32**, 1207–1218 (2011).

58. D. P. Kingma, J. L. Ba, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, (2015).

59. J. Sadr, I. Jarudi, P. Sinha, The role of eyebrows in face recognition, *Perception* (2003).

60. A. Rey, *L'examen clinique en psychologie. [The clinical examination in psychology.]* (1958).

61. W. G. Rosen, Verbal Fluency in Aging and Dementia, *J. Clin. Neuropsychol.* **2**, 135–146 (1980).

62. W. S. Kaplan E, Goodglass H, The Boston Naming Test., *Philadelphia Lea Febiger.* (1983).

63. R. M. REITAN, Validity of the trail making test as an indicator of organic brain damage, *Percept. Mot. Skills* (1958).

64. J. H. Ward, Hierarchical Grouping to Optimize an Objective Function, *J. Am. Stat. Assoc.* **58**, 236–244 (1963).

65. P. Mair, R. Wilcox, Robust statistical methods in R using the WRS2 package, *Behav. Res. Methods* **52**, 464–488 (2020).

66. M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, R. A. Kievit, Raincloud plots: A multi-platform tool for robust data visualization [version 1; peer review: 2 approved], *Wellcome Open Res.* **4** (2019).

Acknowledgments

Funding: Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number R01AG062590. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. 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.

Author contribution: K.K and E.D conceived research; K.K. analyzed data; K.K, K.S.G, M.S and E.D interpreted results. K.K and E.D. wrote the paper.

Competing interest: The authors declare that they have no competing interests.

Data and materials availability: All data associated with this study are in the paper or the Supplementary Materials. All raw data including MRI, CSF, and cognitive scores are available through the ADNI data archive (http://adni.loni.ucs.edu/).

Table 1. Demographics

	AD (A+T+)	**CN (A−T−)**	**MCI**
N	110	109	380
Age	72.91 (8.03)	70.98 (5.32)	72.15 (7.19)
Gender, female	46 (41.82%)	57 (52.29%)	166 (43.68%)
ADAS	21.97 (7.03)	6.97 (3.05)	9.64 (4.43)
CDR	4.58 (1.72)	0.0 (0.07)	1.49 (0.89)
$A\beta_{42}$ (pg/ml)	597.79 (158.44)	1454.75 (247.64)	987.22 (425.97)
p-tau_{181} (pg/ml)	40.41 (14.97)	16.4 (2.82)	27.59 (14.91)

Continuous variables are presented as means, with SDs and categorical variables are presented as % in parentheses. Abbreviations: AD=Alzheimer's disease, CN=cognitively normal, MCI=mild cognitive impairment. N=number of subjects, ADAS=Alzheimer's disease assessment scale, CDR=clinical dementia rating, $A\beta_{42}$=beta-amyloid42, p-tau_{181}=phosphorylated-tau181, SD=standard deviation.

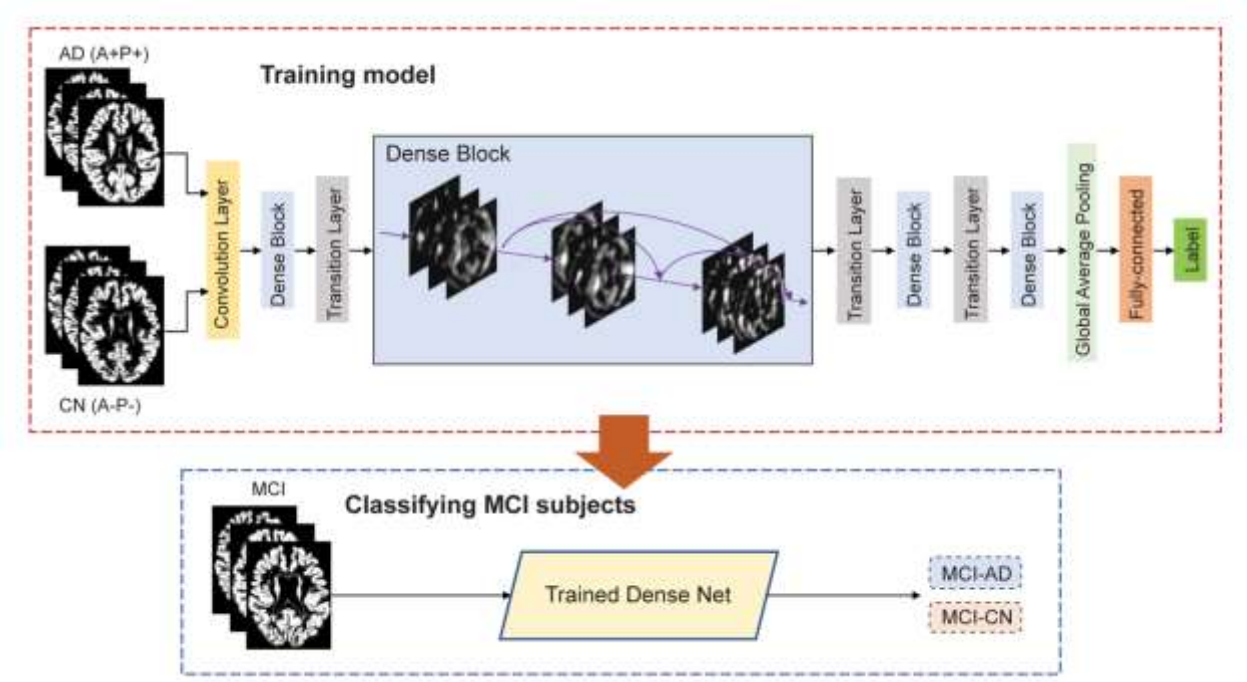

Fig. 1. Study methods. Illustration of proposed deep learning framework. A dense convolutional neural network is trained to differentiate patients with Alzheimer's disease (AD) and cognitively normal (CN) controls based on whole brain gray matter morphometric data. Subsequently, the trained model is deployed to classify individuals with mild cognitive impairment (MCI), into two groups, MCI-AD and MCI-CN based on structural morphometric data.

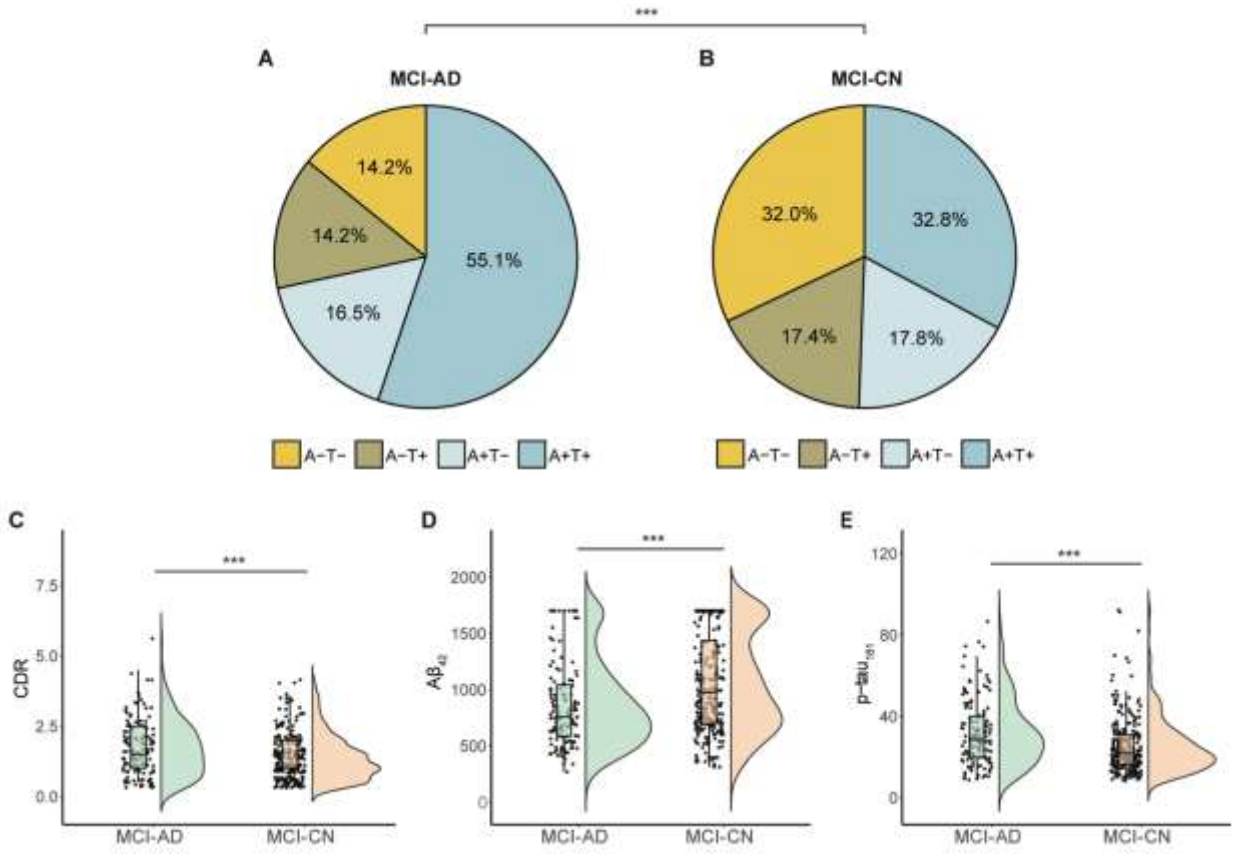

Fig. 2. Validation of the MCI subgroups using fluid biomarker and cognitive data. Subjects in the MCI-AD and MCI-CN groups were rated as Amyloid (A) and p-tau (T), positive or negative, based on the CSF Aβ_{42}, and p-tau$_{181}$ biomarkers. Pie charts depict the biomarker score combinations in the MCI-AD (**A**) and MCI-CN (**B**) subgroups. These score distributions were significantly different between the two subgroups. Box plots show differences in CDR scores (**C**), CSF Aβ_{42} (**D**), and CSF p-tau$_{181}$ (**E**). Abbreviations: AD=Alzheimer's disease, CN=cognitively normal, MCI=mild cognitive impairment, CDR=clinical dementia rating. ***p<0.001.

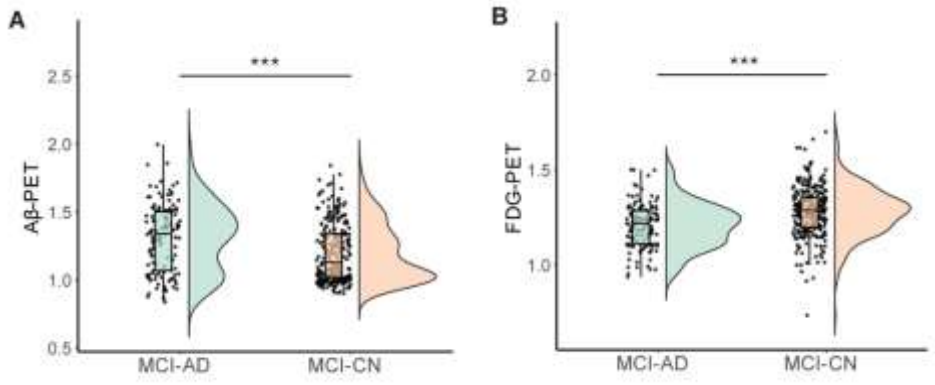

Fig. 3. Comparison of the MCI subgroups using PET uptake data. Boxplots show the comparison of Aβ-PET (**A**) and FDG-PET (**B**) uptake between the MCI-CN and MCI-CD subgroups. In both measurements, group differences were statistically significant. Abbreviations: AD=Alzheimer's disease, CN=cognitively normal, MCI=mild cognitive impairment, Aβ-PET=beta-amyloid positron emission tomography, FDG-PET=fluorodeoxyglucose-positron emission tomography. ***p<0.001

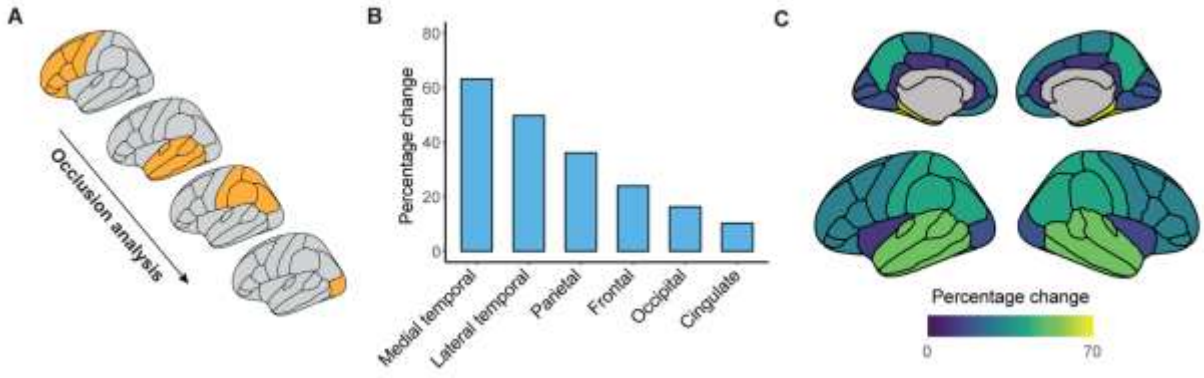

Fig. 4. Identifying the major contributors to atrophy-centered subtyping of MCI via occlusion analysis. (**A**) Schematic illustration of the occlusion analysis. The testing phase in the deep learning model was repeated, whereby in each step, cortical lobes were occluded from the input data (temporal lobes were further divided to medial and lateral). Percentage of change with respect to the original results for classifying MCI subgroups were ranked. (**B**) The results of the occlusion analysis are shown, in each tested lobe. Shown are percentages of change with respect to the original intact model. (**C**) Percentage of change following occlusion analysis, in each cortical lobe, superimposed on medial and lateral cortical surface models. Abbreviations: MCI=mild cognitive impairment

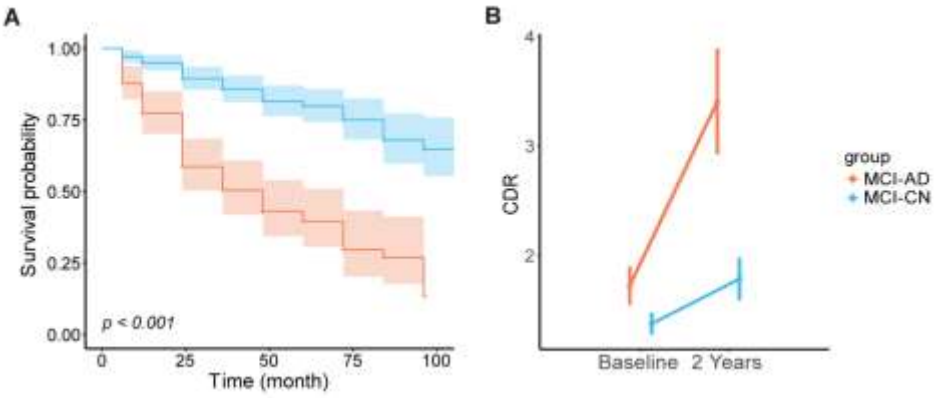

Fig. 5. Longitudinal comparison of the MCI subgroups. (**A**) Kaplan-Meier plots depicting disease-free survival in the MCI-AD and MCI-CN subgroups. The MCI-CN group showed significantly better disease-free survival over time (log-rank test). Shaded areas depict confidence intervals (**B**) Longitudinal changes in CDR scores, displayed by the two MCI subgroups, tested with a RM-ANOVA. Abbreviations: AD=Alzheimer's disease, CN=cognitively normal, MCI=mild cognitive impairment, CDR=clinical dementia rating, RM-ANOVA=repeated measures-analysis of variance.

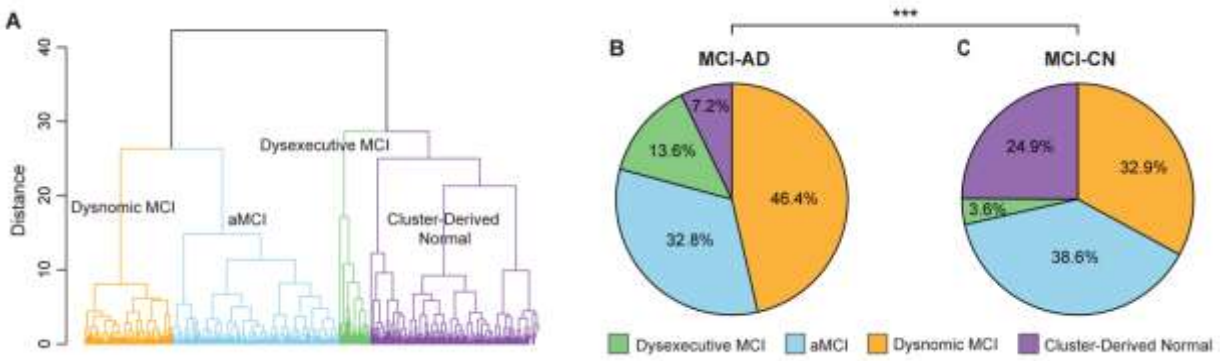

Fig. 6. Overlap between atrophy-centered and neuropsychological subtypes. (**A**) Hierarchical clustering on neuropsychological data was used to define 4 MCI subtypes within the dataset used here (dysnomic MCI, aMCI, dysexecutive MCI, and cluster-derived normal). Pie charts show the prevalence of these subtypes in the MCI-AD (**B**) and MCI-CN (**C**) groups. The two distributions were significantly different. Abbreviations: AD=Alzheimer's disease, CN=cognitively normal, MCI=mild cognitive impairment, aMCI=amnestic mild cognitive impairment. ****p*<0.001