Deep learning identifies brain structures that predict cognition and explain heterogeneity in cognitive aging

Krishnakant V. Saboo[a,b], **Chang Hu**[a,b], **Yogatheesan Varatharajah**[a,b], **Scott A. Przybelski**[b], **Robert I. Reid**[b], **Christopher G. Schwarz**[b], **Jonathan Graff-Radford**[b], **David S. Knopman**[b], **Mary M. Machulda**[b], **Michelle M. Mielke**[b], **Ronald C. Petersen**[b], **Paul M. Arnold**[a,c], **Gregory A. Worrell**[b], **David T. Jones**[b], **Clifford R. Jack Jr**[b], **Ravishankar K. Iyer**[a,1,*], **Prashanthi Vemuri**[b,1,*]

[a]University of Illinois, Urbana-Champaign, IL, United States

[b]Mayo Clinic, Rochester MN, United States

[c]Carle Foundation Hospital, Urbana IL, United States

Abstract

Specific brain structures (gray matter regions and white matter tracts) play a dominant role in determining cognitive decline and explain the heterogeneity in cognitive aging. Identification of these structures is crucial for screening of older adults at risk of cognitive decline. Using deep learning models augmented with a model-interpretation technique on data from 1432 Mayo Clinic Study of Aging participants, we identified a subset of brain structures that were most predictive of individualized cognitive trajectories and indicative of cognitively resilient vs. vulnerable individuals. Specifically, these structures explained why some participants were resilient to the deleterious effects of elevated brain amyloid and poor vascular health. Of these, medial temporal lobe and fornix, reflective of age and pathology-related degeneration, and corpus callosum, reflective of inter-hemispheric disconnection, accounted for 60% of the heterogeneity explained by the most predictive structures. Our results are valuable for identifying cognitively vulnerable individuals and for developing interventions for cognitive decline.

Keywords

Cognitive heterogeneity; Brain reserve; Deep learning; Cognitive aging

1. Introduction

There is considerable heterogeneity in age- and pathology-related cognitive decline in older adults (Stern et al., 2019a; Stern et al., 2019b; Vemuri, 2018; Wilson et al., 2002). While differences in brain health contribute to heterogeneity in cognition (Barulli and Stern, 2013; Satz, 1993; Stern et al., 2019b), the specific brain regions and tracts (along with their interactions) that explain the heterogeneity are not clear (Mungas et al., 2018; Stern et al., 2018; van Loenhoud et al., 2018). In this study, our main objective was to identify brain structures (gray matter regions and white matter tracts) that play a dominant role in the future longitudinal cognitive decline in older adults while quantifying their relative importance in explaining the heterogeneity in cognitive aging. Understanding the protective effect of specific brain structures in mitigating future cognitive consequences of brain pathology has important implications for early diagnosis and for identifying potential targets for stimulation-based therapies (Abellaneda-Perez et al., 2019).

Prior work has limitations because it (i) focused on specific parts of the brain, (ii) did not include aspects that contribute to heterogeneity in the population, or (iii) used simple linear models to capture the complex relationship (Fjell et al., 2013; Nelson et al., 2009) among brain regions, pathology, and cognition (Laubach et al., 2018; Liu et al., 2012; Stern et al., 2018). The hippocampus and the frontal lobe have often been studied (Kaup et al., 2011), especially because of the former's importance in predicting cognitive impairment and Alzheimer's disease, but each of them is specific to one aspect of cognition (i.e., memory (Golomb et al., 1993) and executive function (Zimmerman et al., 2006), respectively). Some studies have evaluated the use of multiple brain regions (Laubach et al., 2018) or white matter tracts (Scott et al., 2017) in predicting cognition and have suggested their utility in studying cognition in aging. However, major factors contributing to the heterogeneity in cognition (e.g., vascular risk) could not be studied because they are not available in typically used datasets (Scott et al., 2017). The relationship between brain structures and cognitive decline can be complex and nonlinear (Nelson et al., 2009). The assumption of a linear relationship, in the above stated and recent machine learning-based studies (Caunca et al., 2021; Kandel et al., 2013; Stonnington et al., 2010), can thus be limiting because it cannot account for complex interactions that contribute to heterogeneity. The preceding considerations are important in studying the role of regions and tracts in cognition in the presence of brain pathologies. We address the above limitations by evaluating several gray matter (GM) regions and white matter (WM) tracts simultaneously in a population-based sample of participants, using a deep learning model to capture the relationship among GM and WM, pathologies, and global cognition.

Accurate modeling and interpretation of the relationship among brain structures, cognition, and pathologies is necessary to identify features that play a dominant role in long-term future cognition in a population. Our approach for doing so consisted of a three-step process:

i. First, we utilized a recently developed deep learning methodology (Saboo et al., 2020) to predict individualized future cognitive trajectories, using baseline GM and WM (i.e., imaging) features, clinical features, and baseline cognition. Deep learning (DL) can extract complex and nonlinear relationships from the data to capture the varying contributions of model features to future cognition across the population and, therefore, can explain more variance in cognition than linear models can. We focused on 25 GM and WM features that span the entire brain and included information about common pathologies that contribute to heterogeneity: amyloid plaque and poor vascular health.

ii. We interpreted the trained deep learning model to quantify the influence of the features on future cognition. We used Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) to do so. For each participant in the study, LIME allowed us to compute the contribution of every feature in the given participant's predicted future cognitive score.

iii. We aggregated the contributions of features (i.e., the LIME value) across participants to study population-level trends in the contributions of various features to future cognition. Population-level trends in LIME values are used to identify imaging features important for cognitive aging because those trends capture the relationships among anatomical features, pathologies, and future cognition that are learned by the DL model.

Our deep learning model was trained on data from a population-based sample of 1432 participants from the Mayo Clinic Study of Aging (MCSA) (Petersen et al., 2010; Roberts et al., 2008). The model predicted individualized global cognitive performance at five time points in the future, i.e., the cognitive trajectory, by using participants' baseline cognition, clinical features (including amyloid and vascular health status), and imaging features (i.e., GM and WM). By interpreting the model using LIME, we identified, in a data-driven manner, seven structures (out of 25) that consistently contributed to the prediction of future cognition in the aging population. The health of those structures contributed to the resilience of some participants to the deleterious effects of elevated brain amyloid and poor vascular health, thus explaining the heterogeneity in cognitive trajectories of those participants. Finally, the overall framework allowed us to quantify the relative importance of those structures in the heterogeneity, enabling a deeper understanding of their role in cognitive aging.

2. Methods

2.1. Study participants

We used participants' data from the Mayo Clinic Study of Aging (MCSA) (Petersen et al., 2010; Roberts et al., 2008) sampled in Olmsted County, Minnesota, USA. Data collection study procedures and ethical aspects were approved by the IRBs of the Mayo Clinic and Olmsted Medical Center and all participants provided an informed consent (Petersen et al., 2010; Roberts et al., 2008). We included participants if (i) we had baseline measurements for them on age, sex, education, genetic risk (*APOE4*), assessment of amyloid pathology, information on vascular health, structural MRI, diffusion tensor imaging (DTI),

and cognitive scores, and (ii) they were at least 50 years old at baseline measurement. Each participant's cognitive scores up to 5 years into the future after baseline were included. Clinical follow-ups in the MCSA dataset are to be 15 months apart, although visit durations of the population are randomly distributed around the 15-month points. To have consistent gaps in age across follow-ups for the entire population, we took a difference between the age at baseline and at follow-up time and rounded that to the nearest integer in terms of years. This was done for every follow-up of a participant and provided the time in years from baseline of a follow-up. Participants who had mild cognitive impairment or dementia based on a consensus of nurse evaluation, risk factor assessment, and neurological and neuropsychological evaluation were classified as "cognitively impaired" (Roberts et al., 2008).

2.2. Imaging assessments

MRI and DTI data were acquired on 3T General Electric (Boston, MA) scanners for all the participants in the study as described in detail in (Vemuri et al., 2018). Our goal was to extract measures of the health of gray and white matter, the two brain tissue classes.

2.2.1. Structural MRI—GM health was measured using cortical thickness (CT). We chose CT measures because it is more closely related to intelligence and cognitive function than volume (see (Menary et al., 2013) for additional support) and reflects the cytoarchitectural characteristics better than volume (Narr et al., 2007). In addition, they are less likely to be influenced by head size (van Loenhoud et al., 2018) and help measure the structural characteristics of the brain, and thus can provide insights into the true neurobiological capital that can be leveraged to understand the impact of brain structures on cognitive performance. CT was computed using an in-house methodology in subregions defined by the Mayo Clinic Adult Lifespan Template (MCALT) ADIR122 Atlas (Schwarz et al., 2016) and can also be found on the NITRC website (https://www.nitrc.org/projects/mcalt/). Cortical thickness values for 82 cortical regions were extracted from the ADIR122 atlas and left and right hemisphere values were averaged to provide 41 subregions (Supplementary Table S1). These were further averaged to retain a smaller subset of ROIs to enable training the deep learning model with the available data.

2.2.2. Diffusion MRI—We used a previously published methodology to process DTI data to obtain FA (Vemuri et al., 2018) in regions defined by the Johns Hopkins University Eve WM Atlas (Oishi et al., 2009). We used fractional anisotropy (FA) instead of mean diffusivity or WM volume because of the greater sensitivity of FA in measuring the microstructural integrity of the WM tracts to determine WM health. Similar to GM regions, we collapsed the 34 sub-tracts into 14 tracts by averaging the FA values of the sub-tracts corresponding to each tract (Vemuri et al., 2018) because we have found that major WM pathways are key for cognitive performance (Raghavan et al., 2021).

2.3. Indicators of pathological processes

Following (Vemuri et al., 2017), we computed a cardiovascular and metabolic conditions (CMC) score by summing the presence or absence of the following conditions within a 5-year period before the baseline visit: hypertension, hyperlipidemia, cardiac arrhythmias,

coronary artery disease, congestive heart failure, diabetes mellitus, and stroke. Using the median CMC of the population, we labeled participants with CMC > 2 as CMC+ and as CMC− otherwise. The presence of amyloid pathology was determined using PET scans as previously described (Jack et al., 2020), using a cut-point of 1.48 SUVR on PiB-PET. We use Aβ+ to indicate those with amyloid above the threshold and Aβ− for the rest. Aβ+/Aβ− and CMC+/CMC− are the two indicators of primary pathological processes that we use in our model.

2.4. Cognitive variables

Our goal for this study was to focus on global cognitive performance. Therefore, the cognition variable was a global measure approximated using a battery of 9 cognitive tests covering the executive function, language, memory, and visuospatial domains (Petersen et al., 2010). The global cognitive z-score was estimated from the average of the z-scores for those 4 domains.

2.5. Model description

The model takes as input baseline cognition score and imaging and clinical features, and outputs the cognition scores over the next five years (Fig. 1A). Specifically, it used the following as inputs: participant's baseline cognition (Cog$_0$), clinical feature (sex, age, education, APOE status, Aβ status, CMC status) and imaging features (CTs and FAs). Modeling of the cognitive trajectory allows the model to capture the influence of brain structure on cognition in the future and the dynamics of this influence in the presence of pathology. Let $x_i^s = [x_{i1}^s, ..., x_{in}^s]$ be the n baseline imaging features and $x_i^c = [x_{i1}^c, ..., x_{im}^c]$ be the m baseline clinical features of participant $i \in \{1, 2, ..., N\}$. Let Cog$_{i0}$ be the baseline cognition score of participant i. **Cog**$_i$ = [Cog$_{i1}$, Cog$_{i2}$, ..., Cog$_{i5}$] corresponds to the cognition scores at future time points at years 1, ..., 5. To explicitly capture the temporal dependency between the cognition scores at two successive time points, we represent the cognition prediction at time t, $\widehat{\text{Cog}}_{it}$, with a function h of the predicted cognition for the previous time point $\widehat{\text{Cog}}_{it-1}$, and a function g_t of the clinical and imaging features:

$$\widehat{\text{Cog}}_{it} = h\big(g_t(r(x_i^s), x_i^c), \widehat{\text{Cog}}_{it-1}\big), t \in \{1, ..., 5\}$$

where $r(x_i^s)$ is a lower-dimensional representation of x_i^s that captures the anatomical information relevant for predicting future cognition. We refer interested readers to (Saboo et al., 2020) for more details.

2.6. Loss function

We used the mean squared error (MSE) and mean absolute error (MAE) for evaluating the model performance. The model was trained to minimize MSE. Since in longitudinal studies, participants might drop out of the study or miss some visits, Cog$_{it}$ might not exist for some participant i at some year t. To make full use of the available data, we modified the MSE loss for N participants to be:

$$\mathscr{L}\left(\left\{\widehat{\mathbf{Cog}}_i, \mathbf{Cog}_i\right\}_{i=1}^N\right) = \frac{\sum_{i=1}^N \sum_{t=1}^T \left(\widehat{\mathrm{Cog}}_{it} - \mathrm{Cog}_{it}\right)^2 \delta_{it}}{\sum_{i=1}^N \sum_{t=1}^T \delta_{it}}$$

where δ_{it} is an indicator function indicating whether Cog_{it} is available in the data. MAE was computed using a similar formula but with the squared error term replaced with absolute error.

2.7. Neural network architecture

We implemented the model using deep neural networks. We chose nonlinear activation for r, g_t and linear activation for h, based on preliminary experiments. r consists of two hidden layers with rectified linear unit (ReLU) activation. g_t constitutes two hidden layers with leaky ReLU activation (negative slope = 0.1). r and g_t were regularized using a dropout. The number of units in r, g_t were varied to give three different architectures: 6-3-3-2, 10-4-4-2, and 12-5-5-2, with the first two numbers providing the units in r and the last two numbers providing the units in g_t. The dropout rate was varied among {0.3, 0.4, 0.5}. We trained all 9 models that resulted from the combinations of architectures and dropout rates by using the same training data, and we evaluated their performance on the validation and test sets. The model (consisting of an architecture and a dropout rate) that achieved the lowest average validation set MSE from cross-validation was used for further analyses (Supplementary Table S2). The best model was 10-4-4-2 with a 0.4 dropout probability.

2.8. Training

80% of the participants were used for training, 10% for validation, and 10% for testing. Baseline features x^s, x^c and baseline cognition Cog_0 were standardized using the training data. Cognition along the trajectory **Cog** was standardized using the standardization parameters of Cog_0. We used the ADAM optimizer (learning rate = 0.001) to train the model end to end, i.e., functions r, g_t, h were trained jointly, for 3000 epochs. From those epochs, we chose the model corresponding to the minimum validation loss.

2.9. Cross-validation

We performed Monte Carlo cross-validation for 50 iterations by randomly splitting the data into 80-10-10 sets in each iteration. Splits were made such that in each iteration, the data for a given participant were present in only one of the sets (i.e., training, validation, or test). We chose a high number of cross-validation iterations to obtain better statistical power. We aggregated model performance and interpretation results on test sets across iterations. For example, we averaged the test loss over the 50 iterations.

2.10. Local Interpretable Model-agnostic Explanations (LIME)

LIME explains the contribution of each feature in a model's output for a sample (Ribeiro et al., 2016). For a given sample, LIME computes a local sparse linear approximation of the trained neural network function on points randomly sampled in the neighborhood of the given sample. Thus, LIME approximates the neural network model locally with a linear model such that the prediction of the linear model and the prediction of the neural

network are the same for the given sample. The sparse linear approximation is leveraged to calculate the contribution of each feature in the prediction for the given sample. A new linear approximation is computed for the explanation of each sample, providing individualized explanations. In our case, LIME provides, for each participant, the real-number contribution of each feature to the person's predicted cognition value. We refer to the contribution as the LIME value. Since for each participant the model predicts cognition at 5 time points, we get separate LIME values for each prediction. For ease of discussion, in the below description of LIME computation, "sample" refers to a given participant and year of prediction.

We computed LIME values using a package made available by its authors (Ribeiro et al., 2016). Since LIME computes a sparse approximation, the package required us to choose the number of features (n') for which the contribution must be computed, with n' being fewer than the number of features input to the model ($n' < n + m + 1$). Therefore, to get reliable LIME values for all the features, we computed LIME values in three steps. In the three steps, we computed LIME values separately for the baseline imaging features, the baseline clinical features (age, sex, education, APOE status, Aβ status, and CMC status), and the baseline cognition (Cog_0), respectively.

Step 1: In computing the LIME values for the 25 imaging features, we set $n' = 10$. In addition, for each sample, we modified all the randomly sampled points to have the same value of the clinical and Cog_0 features as in the given sample. The intuition behind this modification is as follows. LIME relies on the perturbation in the model output to allow the randomly sampled points to compute the feature contributions. If a perturbation in a given feature results in a large change in the model output, then the feature must make a large contribution. Because we set the perturbations for the clinical and Cog_0 features to 0, LIME computes the contribution that results from the imaging features alone. Note that we did not thereby force the contributions of the clinical and Cog_0 features to 0; the contribution of those features is accounted for in the intercept (i.e., constant term) of the linear approximation, since the intercept will be the same for all the randomly sampled points. Thus, by setting $n' = 10$ and setting the perturbation in the clinical and Cog_0 features to zero, we obtained the LIME values for the imaging features for each sample. This provided nonzero contribution values for up to 10 imaging features for the given sample. We set the contribution of the other imaging features to zero for that sample. This choice is justified because if an imaging feature for that sample made a large contribution relative to other imaging features, it would have been in the 10 imaging features that LIME picked as having nonzero contributions. For each sample, LIME also provides the intercept along with the LIME contributions of n' features. To validate the computed LIME values, we summed the intercept with the nonzero LIME values for n' features and 0 contribution from the 25-n' remaining imaging features for the sample. We compared the sum with the prediction of the deep learning model for the given sample. We repeated this comparison for all the samples to evaluate the accuracy of the approximations inherent in LIME as well as those that resulted from the above computational choices (Supplementary Fig. S1A).

Step 2: We repeated the above procedure for computing the LIME value of the clinical features. We set $n' = 7$ to obtain LIME values at least for all the clinical features. Next, for each sample, we set the imaging and Cog_0 features in the randomly sampled points to

have the same value as in the given sample. On applying LIME to that sample, we obtained the contributions of the clinical features. If some clinical feature(s) did not appear in output from the LIME package for a sample, the contribution of those clinical feature(s) for that sample were set to 0. We verified the contributions by summing the intercept and LIME values for the clinical features for the sample and comparing the sum with the prediction from the deep learning model for that sample (Supplementary Fig. S1B).

Step 3: Since h is linear, the relationship between Cog_0 and the predicted cognition is linear and depends only on its coefficient in h (c_0) and the prediction time point t. Therefore, we computed the "LIME-like" value, i.e., the contribution of Cog_0, directly, using h. For each year t of prediction, the contribution of baseline cognition for a sample was set $c_0^t \times Cog_0$.

Finally, to ensure the validity of the contributions computed via the above three procedures, we computed the sum of the LIME values for the imaging features (from step 1), the LIME values for the clinical features (from step 2), and the "LIME-like" value for baseline cognition (from step 3) for each sample, and the intercept of h. We compared the sum with the predicted cognition obtained from the deep learning model (Supplementary Fig. S1C). In all three steps, there was strong agreement between the predicted value of cognition obtained from the model and the sum of the LIME values and intercept (Supplementary Fig. S1). We also evaluated the effect of varying n' for the clinical and imaging features and observed that the LIME values computed for different n' were very similar (Supplementary Fig. S2 and S3). We note that the analyses for identifying top predictors among imaging features rely primarily on LIME values computed in Step 1.

2.11. Clustering to find the top predictors among imaging features

To cluster the features based on their rank for each year in terms of the LIME value, we used the following steps. (i) For each participant, we used the magnitude of his or her LIME value for each feature. (ii) We averaged the absolute LIME value of a feature across participants. (iii) Features were ranked based on their mean absolute LIME value. (iv) Steps (i)–(iii) were repeated separately for each iteration of the cross-validation. (v) Features were hierarchically clustered based on their ranks across the iterations of cross-validation, by means of average linkage hierarchical clustering. (vi) The above procedure from ranking to clustering was repeated separately for each year of prediction. We used ranks based on mean absolute LIME values to find top features because averaging over the absolute LIME values helped retain the contribution for each participant and ranks of the features captured their relative ordering in contribution.

2.12. Relative importance

We were interested in how the health of brain structures influences the differences in cognition in $A\beta+$/CMC+ participants. Since different structures can have varying influences, we measured their relative importance by comparing the extremes (lower vs upper quartile $A\beta+$/CMC+ participants) to capture the range of influence of the health of brain structures on cognition. The relative importance of a feature is a quantification of the difference between upper- and lower-quartile $A\beta+$/CMC+ participants that results from a particular feature, compared to the other top imaging features. To find the participants for computing

relative importance, we followed a three-step procedure. First, for each top predictor, we identified Aβ+/CMC+ participants who had a high (upper-quartile) or low (lower-quartile) feature value (i.e., CT or FA). Second, we took the union of all the participants in the lower quartiles, and the union of all the participants in the upper quartiles. Third, we excluded participants who were in both the sets obtained by taking the union. Thus, we obtained participants who were in the lower quartile or in the upper quartile based on all the top predictors. We used those two groups of participants for computing relative importance (Fig. 5A).

Let S_L be the set of lower-quartile participants and S_U be the set of upper-quartile participants. For a feature $f \in F$ in the set of top predictors, let $\mu_L(f, t)$ be the mean LIME value of f for participants in S_L and $\mu_U(f, t)$ be the mean LIME value of f for participants in S_U for year t of prediction. The relative importance (RI) of the features is computed as follows:

$$RI(f,t) = 100 \times \frac{|\mu_S(f,t) - \mu_L(f,t)|}{\sum_{j \in F} |\mu_S(j,t) - \mu_L(j,t)|} \ \forall f \in F, t \in \{1,...,5\}$$

3. Results

3.1. Data and participants

We used data on 1432 study participants from the MCSA (Petersen et al., 2010; Roberts et al., 2008) sampled in Olmsted County, Minnesota, USA. The data for each participant consisted of baseline (year 0) measurements of cognition, age, sex, education, genetic risk (APOE4), assessment of amyloid pathology, vascular health, MRI, and diffusion tensor imaging (DTI); and cognitive scores up to 5 years into the future after baseline. The number of samples available for each follow-up year were as follows: Year 0 (baseline) – 1432; Year 1 – 595, Year 2 – 252, Year 3 – 419, Year 4 – 561, Year 5 – 521. The participant pool was heterogeneous with respect to the severity of pathologies and cognitive performance (Table 1).

3.2. Performance of cognitive trajectory prediction model

The model predicted individualized cognitive trajectory, i.e., cognition at each year for 5 years after the baseline (Fig. 1). Model performance was measured using mean squared error (MSE) and mean absolute error (MAE). We evaluated several architectures and chose the architecture that achieved the smallest average validation set MSE for further analysis (Supplementary Table S2). We compared the model's test set performance with that of linear regression because linear regression has been widely used for identifying important brain structures in the cognitive aging literature. Our model outperformed linear regression in predicting individualized cognitive trajectories by achieving lower average MSE and MAE across different years of prediction (Wilcoxon signed-rank test, see SI and Supplementary Table S3). These results are consistent with our previous study in which the proposed model outperformed state-of-the-art linear models in 5-year cognition trajectory prediction on ADNI data (Saboo et al., 2020). In (Saboo et al., 2020), improvement in error over linear models was modest for cognitively healthy individuals and substantial for individuals at risk

of future cognitive decline. Modest improvements in prediction in the present study could be because majority of the participants were cognitively unimpaired (Table 1).

3.3.   Individualized and population-level contribution of features to predicted cognition

The trained deep learning model captures the influence of input features on future cognitive scores. Therefore, to identify the brain structures that play an important role in cognitive aging, we interpreted the deep learning model by using a specialized model interpretation methodology called LIME (Ribeiro et al., 2016). For a given participant, LIME provides the contribution of each input feature to the predicted cognitive score at each year for that participant (Fig. 1B). We computed the LIME values for all the participants in the test sets of the cross-validation.

LIME values for the predicted cognitive scores at year 1 for two participants highlight the difference between individuals with respect to the contribution of each feature to their predicted cognition (Fig. 2A and Supplementary Fig. S4). The sum of the LIME values of all the features (plus the model intercept) is consistent with the model's prediction of cognition. Crucially, the LIME values for each feature differ across participants based on the feature's value. For example, as expected, $A\beta$ status contributed to an increase in cognition (LIME value > 0) for participant #1, who was $A\beta-$, but to a decrease in cognition (LIME value < 0) for participant #2, who was $A\beta+$. However, note that the predicted cognition at year 1 was lower than the baseline cognition for both the participants ($-0.06 \rightarrow -0.23$ and $-1.13 \rightarrow -1.44$) because of the combined effect of all the features. That is consistent with the decline seen over time with aging and pathology.

We aggregated the LIME values across participants to explore population-level trends in the contribution of features to cognition (Figs. 1C and 2). We observed that the largest contributors to the prediction of cognition at year 1 were baseline cognition (Cog_0) followed by age. Other prominent contributors were $A\beta$ status, education, whole brain GM (wGM), and whole brain WM (wWM). We also aggregated the LIME values for participants based on their clinical status ($A\beta$ status, CMC status, or sex) (Fig. 2C) and observed that the contributions of those features were consistent with the literature. At a population level, $A\beta+$ status, CMC+ status, as well as male sex contributed to a decrease in cognition. Thus, LIME provides a methodology for interpreting the complex relationships among the features and cognition that were learned by the model, allowing us to explore each feature's contribution to future cognition for a person as well as the feature's role in cognition across the population.

3.4.   The contribution of whole brain GM and WM to cognition

We investigated whether whole brain GM and whole brain WM contribute, on average, to the prediction at a population level after the contributions of all other features have been accounted for. We obtained the contribution of wGM by summing the LIME values of all the GM regions, and the contribution of wWM by summing the LIME values of all the WM tracts, to faithfully capture to cumulative effect of the regions and tracts on predicted cognition. Since different regions and tracts can increase or decrease the predicted cognition, the overall effect of all GM regions and WM tracts would be the cumulative effect across

regions and tracts, respectively. For each year and feature, we averaged the magnitude of the LIME values for all the participants in the test set of each iteration of cross-validation to compute the mean absolute LIME values. Since LIME values can be either positive or negative, using its absolute value allows to retain the contribution for each participant while averaging across the population. The mean absolute LIME values were computed separately for Aβ+ and Aβ– participants (Fig. 2D). Based on these values, Age and Aβ status were the most prominent contributors among the clinical features across five years (Tukey's test, $p < 0.01$ for all pairwise comparisons involving one of Age or Aβ). On average, the contribution of wGM and wWM was smaller for the Aβ– group than the Aβ+ group across the five years of prediction (Wilcoxon rank sum test, $n_1 = n_2 = 50$. $p < 0.001$ for each year). We observed a similar trend while considering mean absolute LIME values stratified based on CMC status (Supplementary Fig. S5). These observations are consistent with previous studies (Jack et al., 2013; Knopman et al., 2013) that have shown that imaging features can influence cognition in the presence of pathology even after the direct effect of the pathological variable has been accounted for.

3.5. Top predictors among the imaging features

Since whole brain GM and WM contribute to cognition on average, we investigated the contributions of each of the 25 imaging features (11 regions and 14 tracts) separately to identify whether some features were consistently more important for predicting cognition than others. LIME values aggregated across all the participants in a test set show that some imaging features (e.g., FA-fornix) make a larger contribution to cognition than others (e.g., CT-temporal pole) in the population (Fig. 3A). Ordering of the imaging features based on their mean absolute LIME values across participants remained consistent over time (Fig. 3B). This suggests that certain imaging features (e.g., FA-fornix and CT-medial temporal lobe) remain consistently important, i.e., have a high mean absolute LIME value (ANOVA, $p < 0.001$ for each year), for prediction of cognition over the years.

To identify the *top predictors* among the imaging features, we clustered the features using their ranks based on their LIME values across different iterations of cross-validation. Since features can make large positive or negative contributions, we used the magnitude of the LIME values to rank the features according to their contribution (see "Methods"). Clustering of imaging features based on their ranks for predicting cognition at year 1 showed that seven features were consistently among the top predictors across iterations: CT-medial temporal lobe (CT-MTL), CT-lateral temporal lobe (CT-LTL), CT-occipital lobe, FA of the fornix (FA-fornix), FA-corpus callosum (FA-CC), FA-cingulum, and FA-internal capsule (Fig. 3C). These seven features were in the cluster of top predictors for all five years (Supplementary Fig. S6), had a high mean absolute LIME value (Fig. 3B), and cumulatively accounted for 64.6% ($\pm 0.2\%$) of the total contribution of all the imaging features across the five years based on the mean absolute LIME value. The consistency of the above seven regions and tracts in having a higher contribution than other features demonstrates that they are important for predicting cognition.

3.6. Differential effect of baseline health of top predictors on cognition in Aβ+/CMC+ participants

From Fig. 2D, we note that wGM and wWM play a role in explaining the heterogeneity in cognitive trajectories between the healthy group (Aβ– or CMC−) and the group with pathological aging (Aβ+ or CMC+). Next, we investigated whether brain structures contributed to explaining the heterogeneity in cognitive aging *within* the "pathology" group. Rephrased differently, does the health of top predictor regions and tracts confer any advantage to participants within the pathology group in their future cognitive trajectories? To investigate this question, we analyzed the effect of cortical thickness and tract integrity on the contribution of the top predictors to cognition in participants who were Aβ+ and CMC+ (i.e., Aβ+/CMC+ participants). In this subset of participants, we classified participants based on whether a given region or tract had a high (upper quartile) or low (lower quartile) CT or FA. We then compared the LIME values for the given region or tract between the upper- and lower-quartile participants for each year (Fig. 4A).

In Aβ+/CMC+ participants, we observed that baseline cortical thickness and tract integrity had a differential effect on cognition for all the top predictors (Fig. 4B). For CT-MTL, CT-LTL, CT-occipital, FA-fornix, FA-cingulum, and FA-CC, a higher CT or FA (corresponding to an upper-quartile participant) made a positive contribution to the prediction of cognition (median LIME > 0), while a lower feature value (for a lower-quartile participant) made a negative contribution (median LIME < 0). This pattern was consistent over the five years of prediction. It suggested that Aβ+/CMC+ participants with a higher CT or FA for those six features at baseline had an advantage in coping with cognitive decline over participants with a lower CT or FA. Strikingly, for FA-internal capsule, we observed the opposite trend. The lower-quartile participants had a positive median LIME value, while the upper-quartile participants had a negative median LIME value. Overall, these observations suggest that the health of the top predictors modified the rate of cognitive decline within the Aβ+/CMC+ participants (Fig. 4C), thus contributing to the heterogeneity within the pathology group.

We performed a confounder analysis to assess whether the top predictors had an independent contribution within Aβ+/CMC+ participants or whether the contributions were simply a result of CT/FA values that reflected the severity of pathology. We observed that each top predictor made an independent contribution to the future cognitive trajectories even after we accounted for the severity of pathology (see SI and Supplementary Fig. S8), supporting the view that those structures contributed to heterogeneity.

3.7. Relative importance of top predictors in explaining heterogeneity

The difference between the upper- and lower-quartile participants varies across the top predictors (Fig. 4B), suggesting that the relative contributions of those brain structures to heterogeneity differ. We quantified the differences to evaluate the relative importance of the top predictors in modifying the cognitive trajectories within the pathology group. We defined the relative importance of a region (or tract) as the change due to that region (or tract) in the upper- and lower-quartile groups of the Aβ+/CMC+ participants relative to the cumulative change from all the top predictors. We were thus able to capture the contribution of each feature to the difference between the two groups within the pathology group.

For that analysis, we identified Aβ+/CMC+ participants who were in the lower-quartile or upper-quartile group for all the top predictors (see "Methods" and Fig. 5A). For these participants, we computed the difference between the mean LIME values of the two groups for each top predictor (Fig. 5B). The relative importance of each feature was the ratio of (i) the magnitude of difference for a feature to (ii) the cumulative magnitude of difference for all the top predictors (see "Methods"). We computed the relative importance separately for each year of prediction (Fig. 5C). For the studied population, we found that FA-fornix had the highest relative importance, followed by CT-MTL and FA-CC. The top three predictors based on relative importance accounted for 61% of the total contribution from all the top predictors. The relative importance of the top predictors was consistent across the years.

3.8. Linear model vs. deep learning model and robustness analyses

We evaluated whether linear regression could have identified the top predictors that we obtained. We used linear regression to predict future cognitive trajectories based on baseline cognition, clinical features, and imaging features, and evaluated the coefficient of each feature and its significance (Supplementary Fig. S7). Like our results, linear regression also identified baseline cognition (Cog_0), age, and Aβ status as significant contributors. Among the imaging features, CT-MTL had significant coefficients for all the years of prediction; FA-fornix had significant coefficients at year 3; and FA-internal capsule had significant coefficients at year 5 prediction. Although the results from linear regression support our findings, the consistency with which the top predictors were found over several years of prediction is unique to our method.

We also assessed the robustness of our approach by training the deep learning model with a modified set of features (baseline cognition, clinical features, and only the seven top predictors) to predict future cognitive trajectories and observed results similar to the ones presented in Figs 3, 4, and 5 (see SI and Supplementary Figs S9, S10, and S11).

4. Discussion

In this discussion, we describe (i) the significance of the results, (ii) the value of the methods used, and (iii) future directions. We leveraged deep learning methodologies to identify brain regions and tracts that are important for long-term cognition and contribute to heterogeneity in aging. Seven features, namely the CT of the medial temporal lobe (CT-MTL), lateral temporal lobe (CT-LTL), and occipital lobe, and the FA of the fornix, corpus callosum (FA-CC), cingulum, and internal capsule, were found to be (i) most predictive of 5-year future cognitive scores in the population; and (ii) critical for modifying the rate of cognitive decline due to pathological processes commonly seen in the elderly (amyloidosis and poor vascular health). Three features contributed to more than 60% of the modification: CT-MTL, FA-fornix, and FA-CC. Our results were obtained via the use of (i) deep learning models that allow for handling of high dimensionality and nonlinearity in data, and (ii) data that capture various factors that contribute to heterogeneity. Our data-driven approach provides empirical evidence for the use of the seven features in future studies to reveal the neural basis for the heterogeneity in cognitive aging and to guide the development of individualized diagnosis techniques and intervention strategies. For example, brain theta-burst stimulation

has been shown to contribute to the maintenance of both cognitive and brain integrity using target stimulation (Abellaneda-Perez et al., 2019; Sole-Padulles et al., 2006).

4.1. Role of MTL, LTL, occipital lobe, corpus callosum, and cingulum

The temporal lobe has been widely used in aging studies because of its roles in supporting cognition and in Alzheimer's disease (AD) (Frisoni et al., 2005; Jack et al., 1992; Scheltens et al., 1992). MTL and LTL were among the three GM regions out of 11 that contributed to prediction of cognition in our model, and that finding validates the use of these regions, specifically MTL, in AD and aging studies (Fjell et al., 2014). The occipital lobe is important for visuospatial processing. Hypoperfusion and atrophy in the occipital lobe has been associated with dementia with Lewy bodies (Hanyu et al., 2006;Prosser et al., 2017), visual hallucinations in Alzheimer's disease (Holroyd et al., 2000), and poor cognitive scores (Smith et al., 2001).

Among the WM measures, a significant amount of research has supported the claim that the corpus callosum and cingulum are important tracts for cognition (Bennett and Madden, 2014; Brickman et al., 2012). The CC comprises the largest interhemispheric connections in the brain and plays a major role in maintaining cognition (Bennett and Madden, 2014). The cingulum bundle has connections to the hippocampus and the limbic system and is strongly involved in the maintenance of cognition (Chao et al., 2013; Kantarci et al., 2011). In a recent ADNI-data-based study of the mean diffusivity of 46 white matter tracts, the CC, cingulum, fornix, and internal capsule were predictive of contemporaneous cognition but not change in cognition (Scott et al., 2017). In contrast, we observed that the WM features were predictive of future cognitive decline in participants with poor vascular health (CMC+) (Supplementary Fig. S5). We were able to make that observation because our study was population-based and included individuals with the complete range of vascular health statuses. Poor vascular health has been shown to be associated with widespread WM injury through several mechanisms, such as increase of white matter hyperintensities and infarctions (van Norden et al., 2012), blood–brain dysfunction (Farrall and Wardlaw, 2009), and chronic inflammation (Jin et al., 2010).

4.2. Fornix as an important predictor and further study of internal capsule

Prior studies have shown the importance of the fornix in predicting cognition in cognitively unimpaired (Fletcher et al., 2013), mildly cognitively impaired (MCI), and AD individuals (Zhuang et al., 2012). While there is evidence for decrease in fornix integrity due to AD-related myelin injury (Desai et al., 2010), the FA-fornix measurements may be highly nonspecific. An important finding of our work is that FA-fornix makes an important contribution, relative to all other brain GM and WM measures, for prediction of cognition, and our work highlights its role in coping with pathologies in the $A\beta$+/CMC+ population (Fig. 4). In line with the conclusion in (Mielke et al., 2012), we found FA-fornix to be as important as CT-MTL for predicting cognition, in terms of relative contribution to cognitive change (Fig. 5). The fornix is connected to the hippocampus, and degradation of the hippocampus-fornix circuit might explain the observed correlation among hippocampal volume, fornix integrity, and cognitive status (Fletcher et al., 2013; Mielke et al., 2012).

Specifically, region-specific loss of fornix integrity is associated with reduced hippocampal volume in MCI and AD patients (Lee et al., 2012).

In a recent study (Scott et al., 2017), the internal capsule was associated with cognition; this supports our finding that the FA-internal capsule is important for predicting future cognition. Among those who are Aβ+/CMC+, lower internal capsule FA contributed to an increase in cognition (median LIME > 0) (Fig. 4). The FA of the internal capsule, though not specific to cognitive performance, may be more sensitive to changes in white matter health as a function of pathology. This finding suggests a compensatory mechanism that has also been observed in a different study for this specific regional white matter structure (Hyett et al., 2018). Further study is warranted to investigate the role of the internal capsule with increasing age-related pathology and its impact on cognition.

While the top predictors among the imaging features we identified have been observed separately in previous studies, and some studies have evaluated either several GM or several WM features for predicting cognition (Laubach et al., 2018; Zhuang et al., 2012), our study utilized several GM and WM features simultaneously. In previous studies, the inclusion of certain imaging features removed the effect of the other imaging features in predicting cognition (Zhuang et al., 2012). An explanation for our observation of the importance of the seven brain structures could be as follows. Brain structures that play a more dominant role in (i.e., make a greater contribution to) future cognition can vary across individuals, e.g., there are participants in our study for whom one region's CT, but not the others', places them in the lower quartile of Aβ+/CMC+ participants. Since deep learning can fit a nonlinear model, it can extract the dominant features for each participant and thus detect several important structures simultaneously in the population. In spite of the differing effects of brain structures on cognition across study participants, a linear model would extract an "average" common effect, which may not reflect the individual-specific underlying relationships among brain regions and cognition in participants.

4.3. Top predictors are suitable proxies for brain reserve and provide resilience to pathology

By modifying the rate of cognitive decline in the Aβ+/CMC+ population, the baseline health of the top predictor brain regions and tracts influenced the relationship between pathology burden and cognitive outcome. Better health (in terms of cortical thickness or fractional anisotropy) for six of those top predictors enhanced a person's ability to cope with cognitive decline (median LIME > 0). If we consider the definition of *brain reserve*— differences in brain size and other quantitative aspects of the brain that explain differential susceptibility to functional impairment in the presence of pathology or other neurological insult (Barulli and Stern, 2013; Stern et al., 2019b)—we could argue that these seven structures fulfil the criteria and are the most eligible brain reserve measures. To assess the utility of the top predictors as brain reserve proxy measures, we also compared their cognition prediction performance with that of commonly used brain reserve proxy measures (see SI). The seven identified features achieved a significantly better performance than the most widely used proxy measure, intracranial volume (Christensen et al., 2009; Groot et al., 2018; van Loenhoud et al., 2018) (Supplementary Table S4). The performance was similar

for CT-MTL alone, indicating that CT-MTL may represent the single most important brain feature that plays a key role and is a critical hub for cognition (Mišić et al., 2014).

While it is highly likely that higher brain reserve provides resilience (i.e., coping ability) against cognitive decline in the presence of pathology (Fig. 4), there are two counterarguments that the identified features are merely reflective of current brain health: (i) the disease may have started earlier in those with lower CT or FA, and it may be a matter of time before we see pathology-related declines in these regions; and (ii) participants with no neurodegeneration can perform better on repeated cognitive tests (Machulda et al., 2017). In confounder analysis, the top predictors contributed to cognition even after the severity of pathology (in the form of Aβ SUVR and CMC) had been accounted for (Supplementary Fig. S8), supporting the view that those brain structures provide resilience against decline and are not merely reflective of disease severity. While one cannot clearly differentiate among the specific mechanisms without substantial longitudinal imaging data, the identification of the brain reserve proxy measures is the first step towards a better mechanistic understanding of the role of brain reserve in cognitive aging.

4.4. Relative importance in explaining heterogeneity

Our methodology allowed us to compute the relative importance of the top predictors among imaging features in the heterogeneity in the Aβ+/CMC+ population. CT-MTL, FA-fornix, and FA-CC were the top 3 and accounted for about 60% of the contribution across the 5 years of prediction. The relative importance of FA-fornix was similar to that of CT-MTL, at 20–25% (Fig. 5) (Fletcher et al., 2013; Mielke et al., 2012). The varying relative importance suggests that several brain structures contribute to the heterogeneity in the population and to varying degrees. Each structure may affect a specific domain of cognition, and the relative importance values could be reflective of cognitive impairment in the studied population (Takeuchi et al., 2017). While the relative importance here was found by comparing the upper and lower quartiles, other measures of relative importance can be defined depending on their utility in downstream analysis. The importance of brain structures may vary across participants, and LIME provides a way of identifying those differences (Fig. 2A). Finally, the importance of imaging features in heterogeneity and individualized diagnosis must be viewed in the context of the larger contributions of clinical features to cognition. To note, the downstream effect of clinical features on cognition are likely mediated through brain regions and tracts.

4.5. Advantage of the employed methodology

We modeled the longitudinal cognitive trajectory based on baseline features with deep learning to model complex relationships and correlations among brain structures, clinical features, and cognition (Stern et al., 2018). Modeling of multiple points on the cognitive trajectory enabled us to identify regions that were consistently contributing to future cognition. The deep learning model achieved better prediction than linear regression, enabling the identification of unique relationships and brain structures not easily obtainable from linear regression (Supplementary Fig. S7). Most importantly, LIME allowed us to "open" the black-box deep learning model and helped in identifying the contribution of

each feature to future cognition in an *individualized* manner (Fig. 2), which may also have translational value for precision medicine (Abellaneda-Perez et al., 2019).

4.6. Limitations and future work

Deep learning models typically require large amounts of data for training. To ensure the replicability of our results, we evaluated the model for many different data splits. Nevertheless, it is necessary to validate the method and results on a larger dataset from other sites. Ideally such a dataset would include information about other aging related pathologies (for e.g., tau and TDP-43 (Jo et al., 2020) as well as scans with multiple protocols. The latter would enable investigating whether the relative importance of the features is maintained when parameters such as voxel size are varied (e.g., for measurement of fornix). LIME computes a local linear approximation of the trained deep learning model in this study and the fidelity of the approximation affects interpretability of the results. While we observed high fidelity approximations (Supplementary Figs. S1, S2, S3), design choices for LIME such as sampling method can influence the approximation. Validating these results with other model interpretation methods such as SHAP (Lundberg and Lee, 2017), DeepLIFT (Avanti et al., 2017), and layer-wise relevance propagation (Dyrba et al., 2021; Montavon et al., 2019) is important.

Our approach, though simple in modelling global cognition as a function of brain health rather than each cognitive domain, provides a broader look, at a coarse level, at the key regions and tracts that impact global cognition. This also allowed us to develop and evaluate the method we proposed, which can now be leveraged to study the relationship between brain structures and specific cognitive domains. Analyses at a finer spatial resolution of subregions and for specific cognitive domains using the proposed approach would further light on the relationships between brain health and cognition in aging.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Abellaneda-Perez K, Vaque-Alcazar L, Vidal-Pineiro D, Jannati A, Solana E, Bargallo N, Santarnecchi E, Pascual-Leone A, Bartres-Faz D, 2019. Age-related differences in default-mode network connectivity in response to intermittent theta-burst stimulation and its relationships with maintained cognition and brain integrity in healthy aging. Neuroimage 188, 794–806. [PubMed: 30472372]

Avanti S, Peyton G, Anshul K, 2017. Learning Important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning. PMLR, Proceedings of Machine Learning Research, pp. 3145–3153 %U.

Barulli D, Stern Y, 2013. Efficiency, capacity, compensation, maintenance, plasticity: emerging concepts in cognitive reserve. Trends Cogn. Sci 17, 502–509. [PubMed: 24018144]

Bennett IJ, Madden DJ, 2014. Disconnected aging: cerebral white matter integrity and age-related differences in cognition. Neuroscience 276, 187–205. [PubMed: 24280637]

Brickman AM, Meier IB, Korgaonkar MS, Provenzano FA, Grieve SM, Siedlecki KL, Wasserman BT, Williams LM, Zimmerman ME, 2012. Testing the white matter retrogenesis hypothesis of cognitive aging. Neurobiol. Aging 33, 1699–1715. [PubMed: 21783280]

Caunca MR, Wang L, Cheung YK, Alperin N, Lee SH, Elkind MSV, Sacco RL, Wright CB, Rundek T, 2021. Machine learning-based estimation of cognitive performance using regional brain MRI markers: the Northern Manhattan Study. Brain Imaging Behav. 15, 1270–1278. [PubMed: 32740887]

Chao LL, Decarli C, Kriger S, Truran D, Zhang Y, Laxamana J, Villeneuve S, Jagust WJ, Sanossian N, Mack WJ, Chui HC, Weiner MW, 2013. Associations between white matter hyperintensities and beta amyloid on integrity of projection, association, and limbic fiber tracts measured with diffusion tensor MRI. PLoS One 8, e65175. [PubMed: 23762308]

Christensen H, Batterham PJ, Mackinnon AJ, Anstey KJ, Wen W, Sachdev PS, 2009. Education, atrophy, and cognitive change in an epidemiological sample in early old age. Am. J. Geriatr. Psychiatry 17, 218–226. [PubMed: 19454848]

Desai MK, Mastrangelo MA, Ryan DA, Sudol KL, Narrow WC, Bowers WJ, 2010. Early oligodendrocyte/myelin pathology in Alzheimer's disease mice constitutes a novel therapeutic target. Am. J. Pathol 177, 1422–1435. [PubMed: 20696774]

Dyrba M, Hanzig M, Altenstein S, Bader S, Ballarini T, Brosseron F, Buerger K, Cantre D, Dechent P, Dobisch L, Duzel E, Ewers M, Fliessbach K, Glanz W, Haynes JD, Heneka MT, Janowitz D, Keles DB, Kilimann I, Laske C, Maier F, Metzger CD, Munk MH, Perneczky R, Peters O, Preis L, Priller J, Rauchmann B, Roy N, Scheffler K, Schneider A, Schott BH, Spottke A, Spruth EJ, Weber MA, Ertl-Wagner B, Wagner M, Wiltfang J, Jessen F, Teipel SJ, Adni A.D.s.g., 2021. Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. Alzheimers Res. Ther 13, 191. [PubMed: 34814936]

Farrall AJ, Wardlaw JM, 2009. Blood-brain barrier: ageing and microvascular disease–systematic review and meta-analysis. Neurobiol. Aging 30, 337–352. [PubMed: 17869382]

Fjell AM, McEvoy L, Holland D, Dale AM, Walhovd KB, 2014. What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus. Prog. Neurobiol 117, 20–40. [PubMed: 24548606]

Fjell AM, Westlye LT, Grydeland H, Amlien I, Espeseth T, Reinvang I, Raz N, Holland D, Dale AM, Walhovd KBAlzheimer Disease Neuroimaging Initiative, 2013. Critical ages in the life course of the adult brain: nonlinear subcortical aging. Neurobiol. Aging 34, 2239–2247. [PubMed: 23643484]

Fletcher E, Raman M, Huebner P, Liu A, Mungas D, Carmichael O, DeCarli C, 2013. Loss of fornix white matter volume as a predictor of cognitive impairment in cognitively normal elderly individuals. JAMA Neurol. 70, 1389–1395. [PubMed: 24018960]

Frisoni GB, Testa C, Sabattoli F, Beltramello A, Soininen H, Laakso MP, 2005. Structural correlates of early and late onset Alzheimer's disease: voxel based morphometric study. J. Neurol. Neurosurg. Psychiatry 76, 112–114. [PubMed: 15608008]

Golomb J, de Leon MJ, Kluger A, George AE, Tarshish C, Ferris SH, 1993. Hippocampal atrophy in normal aging. An association with recent memory impairment. Arch. Neurol 50, 967–973. [PubMed: 8363451]

Groot C, van Loenhoud AC, Barkhof F, van Berckel BNM, Koene T, Teunissen CC, Scheltens P, van der Flier WM, Ossenkoppele R, 2018. Differential effects of cognitive reserve and brain reserve on cognition in Alzheimer disease. Neurology 90, e149–e156. [PubMed: 29237798]

Hanyu H, Shimizu S, Hirao K, Kanetaka H, Sakurai H, Iwamoto T, Koizumi K, Abe K, 2006. Differentiation of dementia with Lewy bodies from Alzheimer's disease using mini-mental state examination and brain perfusion SPECT. J. Neurol. Sci 250, 97–102. [PubMed: 17005204]

Holroyd S, Shepherd ML, Downs JH 3rd, 2000. Occipital atrophy is associated with visual hallucinations in Alzheimer's disease. J. Neuropsychiatry Clin. Neurosci 12, 25–28. [PubMed: 10678508]

Hyett MP, Perry A, Breakspear M, Wen W, Parker GB, 2018. White matter alterations in the internal capsule and psychomotor impairment in melancholic depression. PLoS One 13, e0195672. [PubMed: 29672517]

Jack CR Jr., Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick T, Pankratz VS, Donohue M, Trojanowski JQ, 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12, 207–216. [PubMed: 23332364]

Jack CR Jr., Sharbrough FW, Cascino GD, Hirschorn KA, O'Brien PC, Marsh WR, 1992. Magnetic resonance image-based hippocampal volumetry: correlation with outcome after temporal lobectomy. Ann. Neurol 31, 138–146. [PubMed: 1575452]

Jack CR, Wiste HJ, Weigand SD, Therneau TM, Lowe VJ, Knopman DS, Botha H, Graff-Radford J, Jones DT, Ferman TJ, Boeve BF, Kantarci K, Vemuri P, Mielke MM, Whitwell J, Josephs K, Schwarz CG, Senjem ML, Gunter JL, Petersen RC, 2020. Predicting future rates of tau accumulation on PET. Brain 143, 3136–3150. [PubMed: 33094327]

Jin R, Yang G, Li G, 2010. Inflammatory mechanisms in ischemic stroke: role of inflammatory cells. J. Leukoc. Biol 87, 779–789. [PubMed: 20130219]

Jo M, Lee S, Jeon YM, Kim S, Kwon Y, Kim HJ, 2020. The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. Exp. Mol. Med 52, 1652–1662. [PubMed: 33051572]

Kandel BM, Wolk DA, Gee JC, Avants B, 2013. Predicting cognitive data from medical images using sparse linear regression. Inf. Process. Med. Imaging 23, 86–97. [PubMed: 24683960]

Kantarci K, Senjem ML, Avula R, Zhang B, Samikoglu AR, Weigand SD, Przybelski SA, Edmonson HA, Vemuri P, Knopman DS, Boeve BF, Ivnik RJ, Smith GE, Petersen RC, Jack CR, 2011. Diffusion tensor imaging and cognitive function in older adults with no dementia. Neurology 77, 26–34. [PubMed: 21593440]

Kaup AR, Mirzakhanian H, Jeste DV, Eyler LT, 2011. A review of the brain structure correlates of successful cognitive aging. J. Neuropsychiatry Clin. Neurosci 23, 6–15. [PubMed: 21304134]

Knopman DS, Jack CR, Wiste HJ, Weigand SD, Vemuri P, Lowe VJ, Kantarci K, Gunter JL, Senjem ML, Mielke MM, Roberts RO, Boeve BF, Petersen RC, 2013. Selective worsening of brain injury biomarker abnormalities in cognitively normal elderly persons with β-amyloidosis. JAMA Neurol. 70, 1030–1038. [PubMed: 23797806]

Laubach M, Lammers F, Zacharias N, Feinkohl I, Pischon T, Borchers F, Slooter AJC, Kühn S, Spies C, Winterer G, 2018. Size matters: grey matter brain reserve predicts executive functioning in the elderly. Neuropsychologia 119, 172–181. [PubMed: 30102906]

Lee DY, Fletcher E, Carmichael OT, Singh B, Mungas D, Reed B, Martinez O, Buonocore MH, Persianinova M, Decarli C, 2012. Sub-regional hippocampal injury is associated with fornix degeneration in Alzheimer's disease. Front. Aging Neurosci 4, 1. [PubMed: 22514534]

Liu Y, Julkunen V, Paajanen T, Westman E, Wahlund LO, Aitken A, Sobow T, Mecocci P, Tsolaki M, Vellas B, Muehlboeck S, Spenger C, Lovestone S, Simmons A, Soininen H, 2012. Education increases reserve against Alzheimer's disease–evidence from structural MRI analysis. Neuroradiology 54, 929–938. [PubMed: 22246242]

Lundberg S, Lee SI, 2017. A unified approach to interpreting model predictions. p. arXiv:1705.07874.

Machulda MM, Hagen CE, Wiste HJ, Mielke MM, Knopman DS, Roberts RO, Vemuri P, Lowe VJ, Jack CR Jr., Petersen RC, 2017. Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. Clin. Neuropsychol 31, 99–117. [PubMed: 27724156]

Menary K, Collins PF, Porter JN, Muetzel R, Olson EA, Kumar V, Steinbach M, Lim KO, Luciana M, 2013. Associations between cortical thickness and general intelligence in children, adolescents and young adults. Intelligence 41, 597–606. [PubMed: 24744452]

Mielke MM, Okonkwo OC, Oishi K, Mori S, Tighe S, Miller MI, Ceritoglu C, Brown T, Albert M, Lyketsos CG, 2012. Fornix integrity and hippocampal volume predict memory decline and progression to Alzheimer's disease. Alzheimers Dement 8, 105–113. [PubMed: 22404852]

Miši B, Goñi J, Betzel RF, Sporns O, McIntosh AR, 2014. A network convergence zone in the hippocampus. PLOS Computat. Bio 10, e1003982.

Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R, 2019. Layer-wise relevance propagation: an overview. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing, pp. 193–209 Cham % @ 978-3-030-28954-6.

Mungas D, Gavett B, Fletcher E, Farias ST, DeCarli C, Reed B, 2018. Education amplifies brain atrophy effect on cognitive decline: implications for cognitive reserve. Neurobiol. Aging 68, 142–150. [PubMed: 29798764]

Narr KL, Woods RP, Thompson PM, Szeszko P, Robinson D, Dimtcheva T, Gurbani M, Toga AW, Bilder RM, 2007. Relationships between IQ and regional cortical gray matter thickness in healthy adults. Cereb. Cortex 17, 2163–2171. [PubMed: 17118969]

Nelson PT, Braak H, Markesbery WR, 2009. Neuropathology and cognitive impairment in Alzheimer disease: a complex but coherent relationship. J. Neuropathol. Exp. Neurol 68, 1–14. [PubMed: 19104448]

Oishi K, Faria A, Jiang H, Li X, Akhter K, Zhang J, Hsu JT, Miller MI, van Zijl PC, Albert M, Lyketsos CG, Woods R, Toga AW, Pike GB, Rosa-Neto P, Evans A, Mazziotta J, Mori S, 2009. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants. Neuroimage 46, 486–499. [PubMed: 19385016]

Petersen RC, Roberts RO, Knopman DS, Geda YE, Cha RH, Pankratz VS, Boeve BF, Tangalos EG, Ivnik RJ, Rocca WA, 2010. Prevalence of mild cognitive impairment is higher in men. The Mayo Clinic Study of Aging. Neurology 75, 889–897. [PubMed: 20820000]

Prosser AMJ, Tossici-Bolt L, Kipps CM, 2017. Occipital lobe and posterior cingulate perfusion in the prediction of dementia with Lewy body pathology in a clinical sample. Nucl. Med. Commun 38, 1029–1035. [PubMed: 28926500]

Raghavan S, Reid RI, Przybelski SA, Lesnick TG, Graff-Radford J, Schwarz CG, Knopman DS, Mielke MM, Machulda MM, Petersen RC, Jack CR Jr., Vemuri P, 2021. Diffusion models reveal white matter microstructural changes with ageing, pathology and cognition. Brain Commun. 3, fcab106. [PubMed: 34136811]

Ribeiro MT, Singh S, Guestrin C, 2016. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Roberts RO, Geda YE, Knopman DS, Cha RH, Pankratz VS, Boeve BF, Ivnik RJ, Tangalos EG, Petersen RC, Rocca WA, 2008. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology 30, 58–69. [PubMed: 18259084]

Saboo K, Hu C, Varatharajah Y, Vemuri P, Iyer R, 2020. Predicting longitudinal cognitive scores using baseline imaging and clinical variables. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1326–1330.

Satz P, 1993. Brain reserve capacity on symptom onset after brain injury: a formulation and review of evidence for threshold theory. Neuropsychology 7, 273.

Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, Kuiper M, Steinling M, Wolters EC, Valk J, 1992. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J. Neurol. Neurosurg. Psychiatry 55, 967–972. [PubMed: 1431963]

Schwarz CG, Gunter JL, Wiste HJ, Przybelski SA, Weigand SD, Ward CP, Senjem ML, Vemuri P, Murray ME, Dickson DW, Parisi JE, Kantarci K, Weiner MW, Petersen RC, Jack CR Jr., 2016. A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. Neuroimage Clin. 11, 802–812. [PubMed: 28050342]

Scott JA, Tosun D, Braskie MN, Maillard P, Thompson PM, Weiner M, DeCarli C, Carmichael OT, 2017. Independent value added by diffusion MRI for prediction of cognitive function in older adults. Neuroimage Clin. 14, 166–173. [PubMed: 28180075]

Smith MZ, Esiri MM, Barnetson L, King E, Nagy Z, 2001. Constructional apraxia in Alzheimer's disease: association with occipital lobe pathology and accelerated cognitive decline. Dement. Geriatr. Cogn. Disord 12, 281–288. [PubMed: 11351139]

Sole-Padulles C, Bartres-Faz D, Junque C, Clemente IC, Molinuevo JL, Bargallo N, Sanchez-Aldeguer J, Bosch B, Falcon C, Valls-Sole J, 2006. Repetitive transcranial magnetic stimulation effects on brain function and cognition among elders with memory dysfunction. A randomized sham-controlled study. Cereb. Cortex 16, 1487–1493. [PubMed: 16339086]

Stern Y, Arenaza-Urquijo EM, Bartres-Faz D, Belleville S, Cantilon M, Chetelat G, Ewers M, Franzmeier N, Kempermann G, Kremen WS, Okonkwo O, Scarmeas N, Soldan A, Udeh-Momoh C, Valenzuela M, Vemuri P, Vuoksimaa E, 2018. Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimers Dement.

Stern Y, Barnes CA, Grady C, Jones RN, Raz N, 2019a. Brain reserve, cognitive reserve, compensation, and maintenance: operationalization, validity, and mechanisms of cognitive resilience. Neurobiol. Aging 83, 124–129. [PubMed: 31732015]

Stern Y, Chételat G, Habeck C, Arenaza-Urquijo EM, Vemuri P, Estanga A, Bartrés–Faz D, Cantillon M, Clouston SAP, Elman JA, Gold BT, Jones R, Kempermann G, Lim YY, van Loenhoud A, Martínez-Lage P, Morbelli S, Okonkwo O, Ossenkoppele R, Pettigrew C, Rosen AC, Scarmeas N, Soldan A, Udeh-Momoh C, Valenzuela M, Vuoksimaa E, 2019b. Mechanisms underlying resilience in ageing. Nat. Rev. Neurosci 20, 246. [PubMed: 30814677]

Stonnington CM, Chu C, Kloppel S, Jack CR, Ashburner J, Frackowiak RSAlzheimer Disease Neuroimaging, Initiative, 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51, 1405–1413. [PubMed: 20347044]

Takeuchi H, Taki Y, Nouchi R, Yokoyama R, Kotozaki Y, Nakagawa S, Sekiguchi A, Iizuka K, Yamamoto Y, Hanawa S, Araki T, Miyauchi CM, Shinada T, Sakaki K, Sassa Y, Nozawa T, Ikeda S, Yokota S, Daniele M, Kawashima R, 2017. Global associations between regional gray matter volume and diverse complex cognitive functions: evidence from a large sample study. Sci. Rep 7, 10014. [PubMed: 28855703]

van Loenhoud AC, Groot C, Vogel JW, van der Flier WM, Ossenkoppele R, 2018. Is intracranial volume a suitable proxy for brain reserve? Alzheimers Res. Ther 10, 91. [PubMed: 30205838]

van Norden AG, van Dijk EJ, de Laat KF, Scheltens P, Olderikkert MG, de Leeuw FE, 2012. Dementia: Alzheimer pathology and vascular factors: from mutually exclusive to interaction. Biochim. Biophys. Acta 1822, 340–349. [PubMed: 21777675]

Vemuri P, 2018. Exceptional brain aging without Alzheimer's disease: triggers, accelerators, and the net sum game. Alzheimers Res. Ther 10, 53. [PubMed: 29859131]

Vemuri P, Lesnick TG, Przybelski SA, Graff-Radford J, Reid RI, Lowe VJ, Zuk SM, Senjem ML, Schwarz CG, Gunter JL, Kantarci K, Machulda MM, Mielke MM, Petersen RC, Knopman DS, Jack CR Jr., 2018. Development of a cerebrovascular magnetic resonance imaging biomarker for cognitive aging. Ann. Neurol 84, 705–716. [PubMed: 30264411]

Vemuri P, Lesnick TG, Przybelski SA, Knopman DS, Lowe VJ, Graff-Radford J, Roberts RO, Mielke MM, Machulda MM, Petersen RC, Jack CR, 2017. Age, vascular health, and Alzheimer disease biomarkers in an elderly sample. Ann. Neurol 82, 706–718. [PubMed: 29023983]

Wilson RS, Beckett LA, Barnes LL, Schneider JA, Bach J, Evans DA, Bennett DA, 2002. Individual differences in rates of change in cognitive abilities of older persons. Psychol. Aging 17, 179–193. [PubMed: 12061405]

Zhuang L, Sachdev PS, Trollor JN, Kochan NA, Reppermund S, Brodaty H, Wen W, 2012. Microstructural white matter changes in cognitively normal individuals at risk of amnestic MCI. Neurology 79, 748–754. [PubMed: 22843270]

Zimmerman ME, Brickman AM, Paul RH, Grieve SM, Tate DF, Gunstad J, Cohen RA, Aloia MS, Williams LM, Clark CR, Whitford TJ, Gordon E, 2006. The relationship between frontal gray matter volume and cognition varies across the healthy adult lifespan. Am. J. Geriatr. Psychiatry 14, 823–833. [PubMed: 17001022]
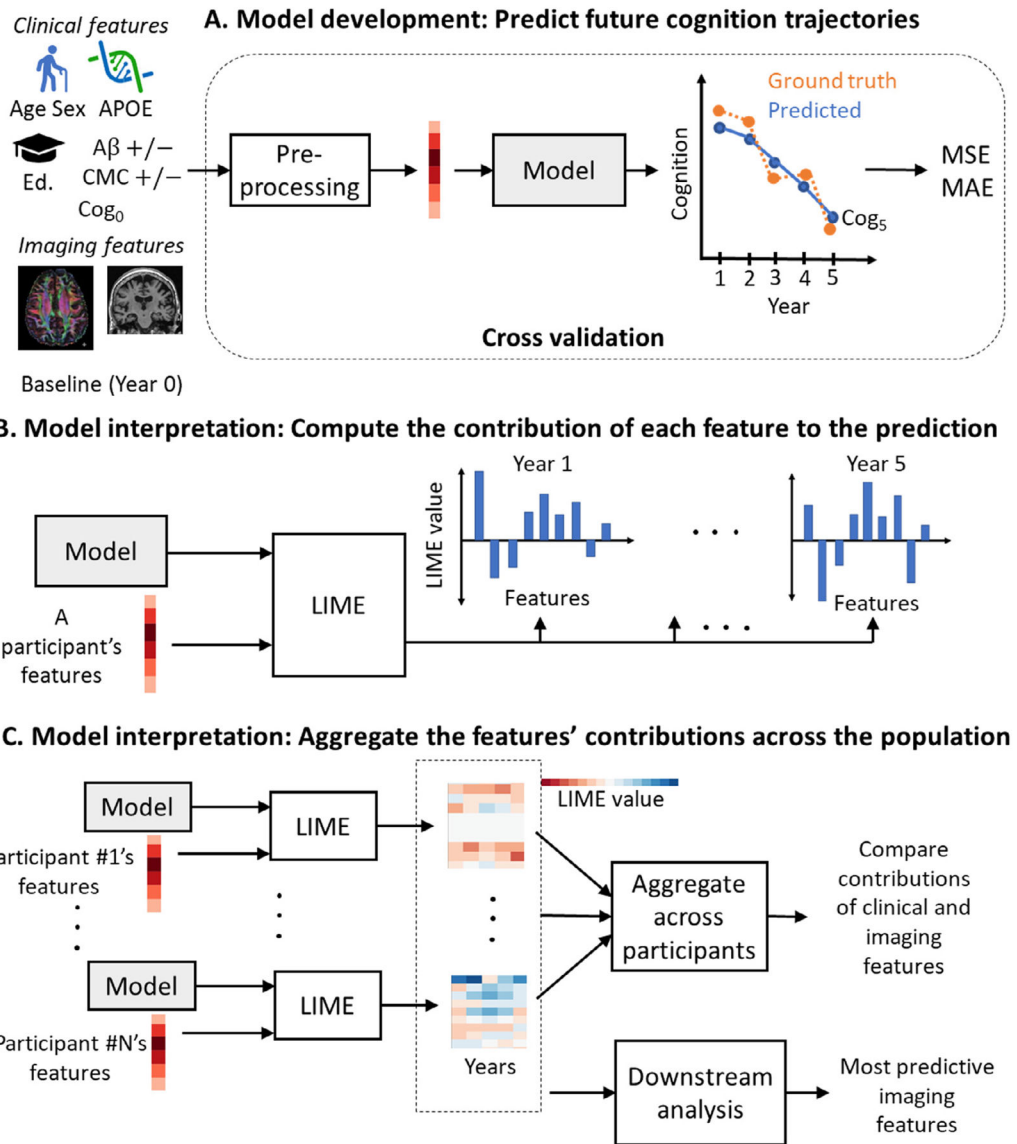
Fig. 1.

Overall analysis workflow for identifying brain regions and tracts important for long-term cognition in the aging population. (A) Model development: Deep learning model for predicting the 5-year future cognitive trajectory of a participant based on baseline imaging features (GM and WM), clinical features (age, sex, education, APOE4 genotype, amyloid status, and systemic vascular health), and baseline cognition (Cog_0). Cognition is measured using global cognitive performance. Model predictions are compared with ground truth to evaluate model performance. (B) Model interpretation: Local Interpretable Model-agnostic Explanations (LIME) makes it possible to interpret the trained deep learning model by determining the contribution of a participant's features to the predicted cognition for a given participant. Explanations (LIME values) are obtained for each year of prediction for a participant. (C) Model interpretation: LIME values are obtained for several participants in the population for each year and aggregated to compare the contributions of clinical

and imaging features. Participants' LIME values are used for further analysis to identify important brain regions and tracts. Abbreviations: Aβ, amyloid; CMC, cardiovascular and metabolic conditions; Ed., education; MSE, mean squared error; MAE, mean absolute error.
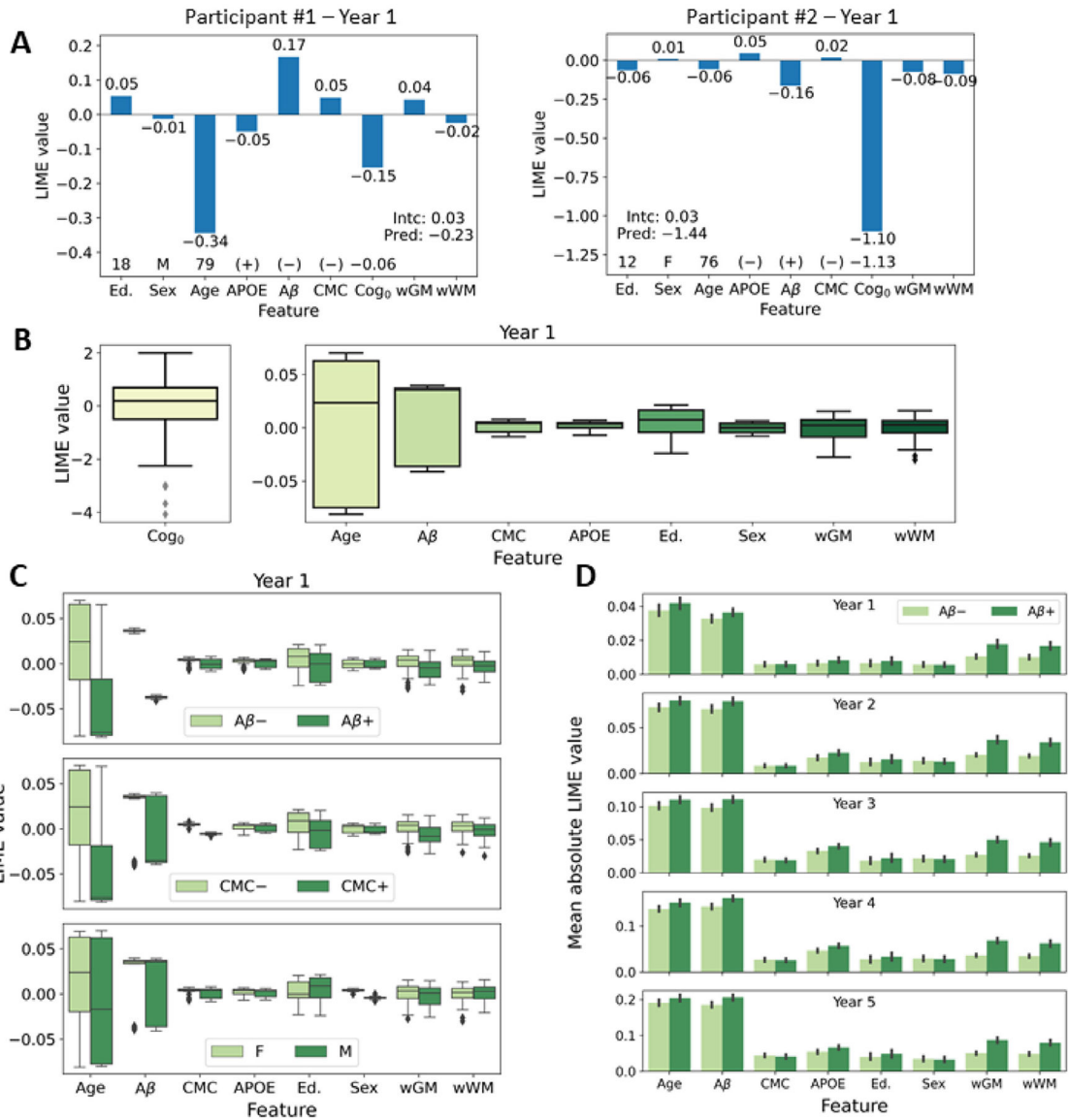
Fig. 2.
Contribution of the features to predicted cognition obtained by interpreting the deep learning model using LIME. (A) LIME value of baseline features for predicting cognition at year 1 for two different participants. Feature values for the participants are provided at the bottom. The sum of the LIME values of all the features and the intercept ("Intc") is equal to the predicted cognitive score ("Pred"). LIME for wGM was obtained by summing the LIME values of all the GM regions, and for wWM by summing the LIME values of all the WM regions. (B) LIME values for year 1 predictions for all the participants in the test set for one iteration in the cross-validation. LIME values for baseline cognition (Cog_0) are provided separately for visual clarity. (C) LIME values for year 1 for one iteration grouped based on the status of three clinical variables (amyloid levels, systemic vascular health, and sex). LIME values for cognition are not shown for visual clarity. (D) Mean absolute LIME values over the cross-validation for each year of prediction stratified by $A\beta$ status. Mean

absolute LIME value of a feature for Aβ+ for an iteration of cross-validation is computed by averaging the magnitude of the LIME values of the given feature for all Aβ+ participants in the test set of the given iteration. Abbreviations: Ed, education; wGM, whole brain gray matter; wWM, whole brain white matter.
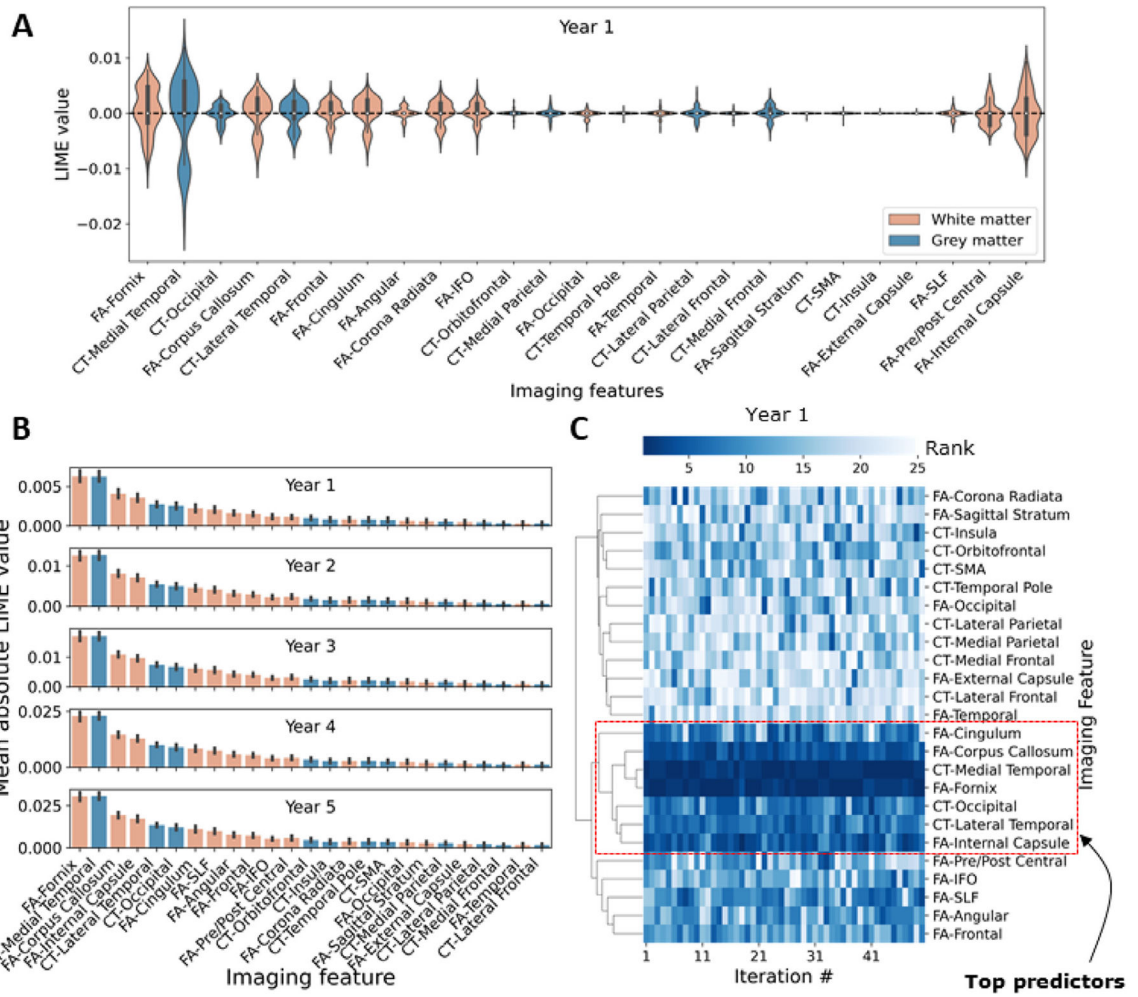
Fig. 3.

Contribution of the imaging features to prediction of future cognition. (A) Violin plot of LIME values for all the imaging features for year 1 prediction for the participants in one iteration. (B) Mean absolute LIME values across cross-validation for each year of prediction for the imaging features. (C) Hierarchical clustering of imaging features based on their ranks in different iterations of cross-validation for year 1 prediction. Ranks were computed based on the mean absolute LIME value. The top predictors among imaging features were obtained by considering the cluster of the best-ranked imaging features across iterations. Abbreviations: CT, cortical thickness; FA, fractional anisotropy; IFO, inferior fronto-occipital fasciculus; SMA, supplementary motor area; SLF, superior longitudinal fasciculus.
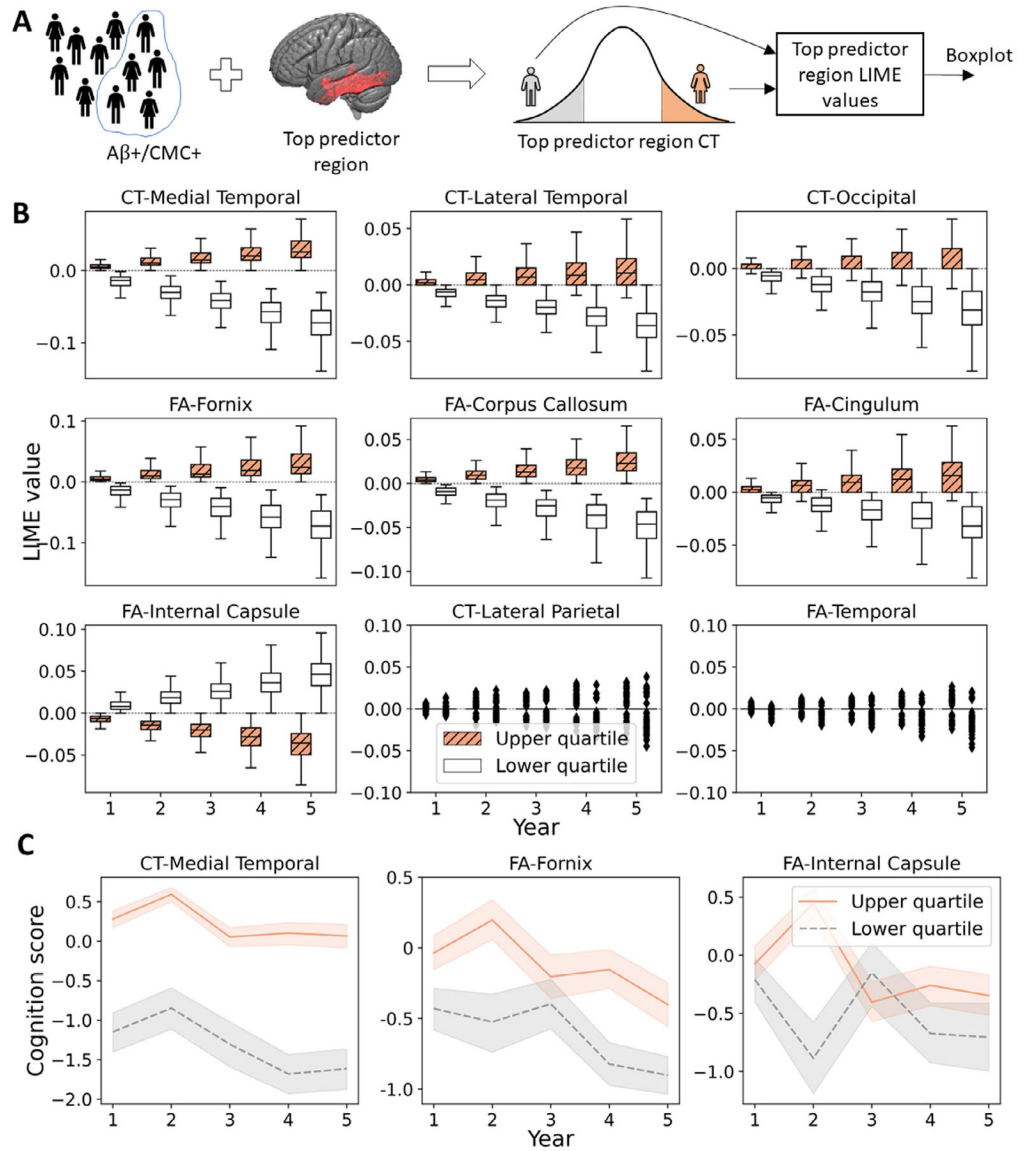
Fig. 4.
Top predictors among imaging features contribute to the heterogeneity within the pathology group. (A) Participants with elevated brain amyloid (Aβ+) and poor vascular health (CMC+) (Aβ+/CMC+) are considered. For each top predictor region, participants with high (upper quartile) or low (lower quartile) CT of the given region are identified. LIME values of the given region for upper- and lower-quartile participants are further analyzed. The same is done for top predictor tracts. (B) Boxplot of LIME values over time for upper- and lower-quartile Aβ+/CMC+ participants. Baseline health of top predictor regions and tracts modifies the rate of cognitive decline in the pathology group by contributing to an increase or decrease in cognition. Plots are shown for all seven top predictors and for two imaging features that were not identified as top predictors (CT-Lateral Parietal and FA-Temporal). (C) The average predicted future cognitive trajectories for upper- and lower-quartile participants for three top predictors.
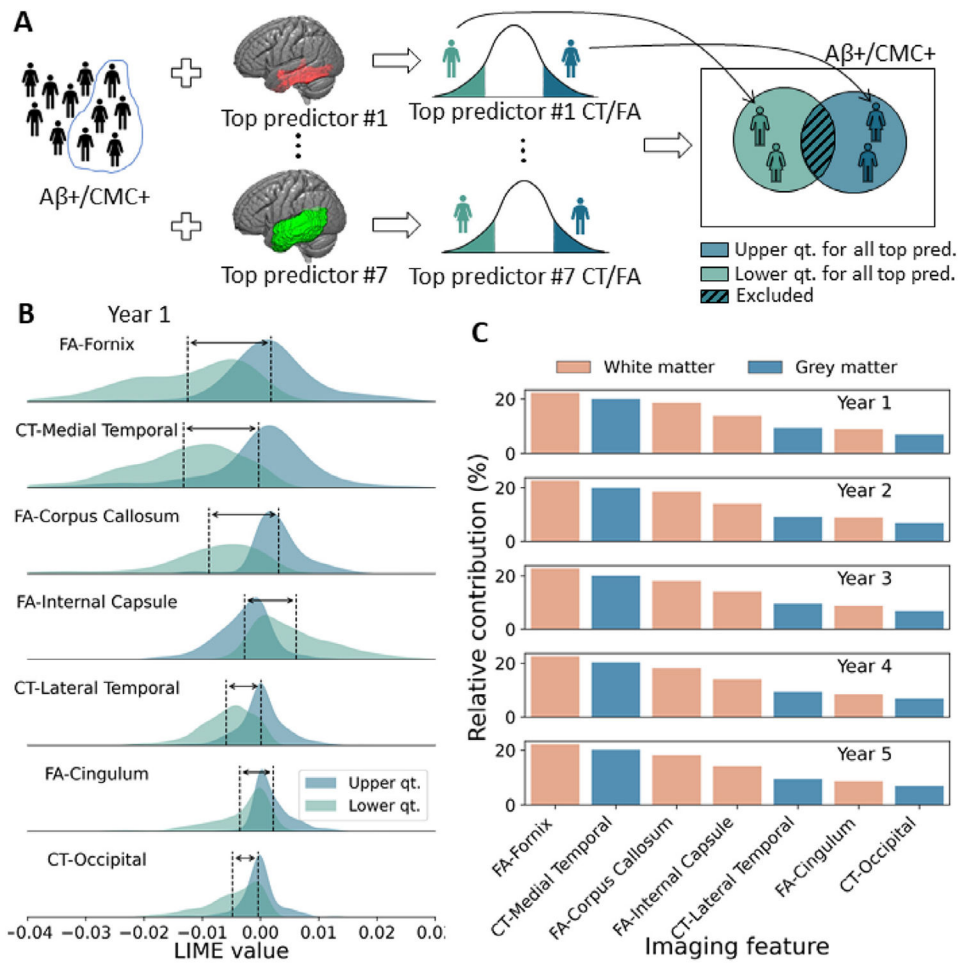
Fig. 5.

Relative importance of top predictors among imaging features in modifying the cognitive trajectories within the pathology group. (A) Identification of Aβ+/CMC+ participants who were in the upper quartile or lower quartile for all top predictors based on the CT (or FA) of those regions (or tracts). (B) Distribution of each top predictor's LIME values for year 1 prediction for the upper- and lower-quartile participants. The difference between the mean LIME values (dashed vertical lines) of the two groups was obtained. The difference for a feature represents the average modification in the cognitive trajectory in Aβ+/CMC+ participants that resulted from that feature. (C) The relative importance of the top predictors for each year of prediction. Relative importance is the percentage of difference caused by a feature compared to the cumulative difference caused by all the features (i.e., all top predictors).

Table 1

Participants' characteristics table with the mean (SD) listed for continuous variables and count (%) for categorical variables. Cognitive diagnosis information was unavailable for 3 individuals in the 50-59 age group, 1 in the 60-69 group, 1 in the 70-79 group, and 0 in the 80-99 group. The square brackets indicate the number of missing values for each age group.

	50–59 n = 250	60–69 n = 454	70–79 n = 452	80–99 n = 276
CLINICAL FEATURES				
Age, yrs	55.0 (2.5)	64.6 (2.8)	74.5 (2.8)	83.9 (3.3)
Males	130 (52%)	231 (51%)	242 (54%)	169 (61%)
APOE4 Carrier	73 (29%)	133 (29%)	137 (30%)	67 (24%)
Education	15.1 (2.2)	15.1 (2.4)	14.5 (2.6)	14.6 (3.1)
Vascular Health (CMC+)	29 (12%)	119 (26%)	199 (44%)	162 (59%)
Amyloid Positive ($A\beta$+)	6 (2%)	80 (18%)	159 (35%)	159 (58%)
COGNITION-RELATED VARIABLES				
Baseline Global Cognition	0.86 (0.82)	0.44 (0.87)	−0.15 (1.0)	−0.64 (1.2)
Total Cognitive Follow-ups	0.8 (1.6)	1.6 (1.9)	2.2 (1.8)	1.6 (1.8)
Cognitively Impaired [3, 1, 1, 0]	4 (2%)	24 (5%)	49 (11%)	55 (20%)