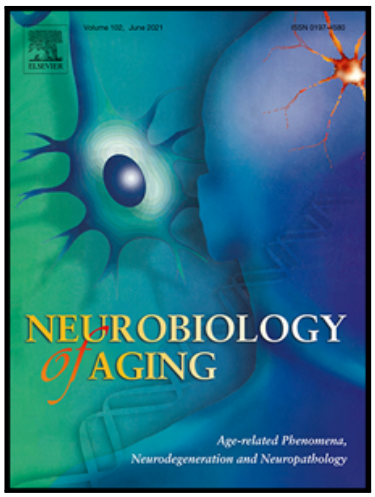

Predicting future cognitive decline from non-brain and multimodal brain imaging data in healthy and pathological aging

Bruno Hebling Vieira , Franziskus Liem , Kamalaker Dadi , Denis A. Engemann , Alexandre Gramfort , Pierre Bellec , R. Cameron Craddock , Jessica S. Damoiseaux , Christopher J. Steele , Tal Yarkoni , Nicolas Langer , Daniel S. Margulies , Gaël Varoquaux

PII: S0197-4580(22)00140-3
DOI: https://doi.org/10.1016/j.neurobiolaging.2022.06.008
Reference: NBA 11386

To appear in: *Neurobiology of Aging*

Received date: 11 August 2020
Revised date: 21 June 2022
Accepted date: 23 June 2022

Please cite this article as: Bruno Hebling Vieira , Franziskus Liem , Kamalaker Dadi , Denis A. Engemann , Alexandre Gramfort , Pierre Bellec , R. Cameron Craddock , Jessica S. Damoiseaux , Christopher J. Steele , Tal Yarkoni , Nicolas Langer , Daniel S. Margulies , Gaël Varoquaux , Predicting future cognitive decline from non-brain and multimodal brain imaging data in healthy and pathological aging, *Neurobiology of Aging* (2022), doi: https://doi.org/10.1016/j.neurobiolaging.2022.06.008

Predicting future cognitive decline from non-brain and multimodal brain imaging data in healthy and pathological aging

Bruno Hebling Vieira[1,2,#], Franziskus Liem[3,#], Kamalaker Dadi[4], Denis A. Engemann[4,5], Alexandre Gramfort[4], Pierre Bellec[6], R. Cameron Craddock[7], Jessica S. Damoiseaux[8], Christopher J. Steele[9], Tal Yarkoni[7], Nicolas Langer[1,2,3], Daniel S. Margulies[10], Gaël Varoquaux[4]

[1]Methods of Plasticity Research, Department of Psychology, University of Zurich, Zurich, Switzerland

[2]Neuroscience Center Zurich (ZNZ), University of Zurich & ETH Zurich, Zurich, Switzerland

[3]University Research Priority Program "Dynamics of Healthy Aging", University of Zurich, Zurich, Switzerland

[4]Université Paris-Saclay, Inria, CEA, Palaiseau, France

[5]Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[6]University of Montreal, Montreal, Canada

[7]The University of Texas, Austin, TX, USA

[8]Wayne State University, Detroit, MI, USA

[9]Concordia University, Montreal, Canada

[10]Institut du Cerveau et de la Moelle épinière, Paris, France

[#]Equal contribution, the order was determined alphabetically

Highlights

- We tested if multimodal neuroimaging improves the prediction of future cognition
- Adding structural data improved the continuous prediction of rate of decline
- Best test-set performance in healthy and pathological aging: $R^2 = 0.42$
- Cognitive performance, daily functioning, and subcortical volume drove performance
- In contrast, including functional connectivity did not improve predictive accuracy

Abstract

Previous literature has focused on predicting a diagnostic label from structural brain imaging. Since subtle changes in the brain precede cognitive decline in healthy and pathological aging, our study predicts future decline as a continuous trajectory instead. Here, we tested whether baseline multimodal neuroimaging data improve the prediction of future cognitive decline in healthy and pathological aging. Non-brain data (demographics, clinical and neuropsychological scores), structural MRI and functional connectivity data from OASIS-3 (N=662; age=46–96y) were entered into cross-validated multi-target random forest models to predict future cognitive decline (measured by CDR and MMSE), on average 5.8y into the future. The analysis was preregistered, and all analysis code is publicly available. Combining non-brain with structural data improved the continuous prediction of future cognitive decline (best test-set performance: R_2=0.42). Cognitive performance, daily functioning, and subcortical volume drove the performance of our model. Including functional connectivity did not improve predictive accuracy. In the future, the prognosis of age-related cognitive decline may enable earlier and more effective individualized cognitive, pharmacological, and behavioral interventions.

Keywords

biomarker, machine learning, predictive modeling, cross-validation, open science

1 Introduction

Cognitive decline, such as worsening memory or executive functioning, occurs in healthy and pathological aging. Crucially, noticeable decline may be preceded by subtle changes in the brain. It is this sequence that enables using brain imaging data to predict the current cognitive functioning of a person or related surrogate markers. For example, structural brain imaging has been used to predict patients' current cognitive diagnosis (Rathore et al. 2017), or brain-age (Cole and Franke 2017), a surrogate biomarker related to cognitive impairment (Liem et al. 2017). Together, these findings demonstrate the clinical potential of neuroimaging data used in combination with predictive analyses.

While predicting *current* cognitive functioning enables insight into related brain markers, predicting *future* cognitive decline from baseline data poses a greater challenge with more substantial clinical relevance (Davatzikos 2019). Using current brain imaging data to predict a current diagnostic label (such as dementia), targets a label that can fairly easily be determined via other means such as clinical assessments (and usually with less cost than brain imaging). When predicting future cognitive change, however, brain imaging might aid a prognosis with greater clinical utility that cannot be easily obtained otherwise. Most previous studies that predicted future change restricted their analysis to whether patients with mild cognitive impairment (MCI) converted to Alzheimer's disease (AD) (e.g., Eskildsen et al. 2015; Korolev et al. 2016; Gaser et al. 2013; Davatzikos et al. 2011) or predicted membership in data-driven trajectory-groups of future decline (Bhagwat et al. 2018). See Table 1 for a comparison. Predicting future cognitive decline on a continuum (instead of forming distinct diagnostic labels from cognitive data) better characterizes the underlying change in abilities on an individual level. This approach can also be used to widen the scope

of applications by including healthy aging. Brain data is a rich source of information that might help us better understand and even reorganize diagnostic syndromes or categories.

Table 1 Comparison of a non-exhaustive selection of studies that perform prediction of future cognitive decline in the context of AD. Compared with the literature, which is often concerned with predicting discrete class assignment, our approach predicts rates of cognitive decline based on a wide selection of input features.

	Targets	Inputs	Analysis method
(Eskildsen et al. 2015)		Regional cortical thickness, non-local hippocampal morphological grading scores, clinical scores (MMSE, RAVLT), age	Linear discriminant analysis with multivariate feature selection
(Korolev et al. 2016)	MCI to AD conversion	Clinical scores (risk factors, clinical assessments, medication status), regional GM morphometry, (cortical and subcortical volumes, mean cortical thickness, standard deviation of cortical thickness, surface area, curvature), plasma proteomics biomarkers	Probabilistic multiple kernel learning (pMKL) with multivariate feature selection
(Gaser et al. 2013)		Estimated “brain age” score, baseline clinical scores, age, hippocampal volume	Cox regression, ROC analysis
(Davatzikos et al. 2011)		Automated marker of atrophy (SPARE-AD), CSF biomarkers	SVM

	Targets	Inputs	Analysis method
(Bhagwat et al. 2018)	Membership to clusters of MMSE and ADAS-13 trajectories	Regional cortical thickness, APOE4 status, age and baseline clinical scores	Longitudinal siamese network
Current study	Rate of MMSE and CDR-SOB change	Mean cortical thickness, GM, WM and CSF total volumes, regional subcortical volumes, functional connectivity, age, APOE status, baseline clinical scores, demographic information, health status, neuropsychological assessment	Multi-target Random Forest

While most previous predictive studies used structural brain imaging alone, integrating structural and functional imaging has been shown to improve predictions. Since both brain structure (Oschwald et al. 2019) and brain function (Liem et al. 2020) change in aging, the most accurate predictions of brain-age have come from combining them (Liem et al. 2017; Engemann et al. 2020; Schulz et al. 2022). Multimodal gains have also been shown in more complex predictions such as current diagnosis in AD (Rahim et al. 2016) and conversion from MCI to AD (e.g., Hojjati et al. 2018; Dansereau et al. 2017; Tam et al. 2019). Therefore, integrating multiple brain imaging modalities enables a more complete characterization of brain aging and provides increased predictive power.

Demographic, health and clinical variables, which are straightforward to obtain and non-invasive, were demonstrated to reliably predict the conversion to cognitive impairment in elders over a 2-year period (Na 2019) in the absence of any brain imaging data, demonstrating the value of non-brain data. Likewise, non-brain data pertaining to mood, demographics, lifestyle, education and early-life factors were shown to be on par with brain

imaging data in the prediction intelligence and neuroticism, but not brain-age delta (Dadi et al. 2021).

The present study aimed to predict future cognitive decline from baseline data in healthy and pathological aging. We combined non-brain data, such as scores from clinical assessments and demographics, with multimodal brain imaging data to test whether adding brain imaging to non-brain data improves predictive performance, and whether multimodal imaging outperforms single imaging modalities. We showed that structural imaging in particular improved continuous prediction of future cognitive decline. An early prognosis of future cognitive decline might enable earlier and more effective pharmacological or behavioral treatments to be tailored to the individual, resulting in more efficiently allocated medical resources.

2 Methods

The analysis presented here was preregistered (Liem et al. 2019). We largely followed this preregistration and deviations are described in the supplement (*6.1.2 Deviation from preregistration*). The deviations concern minor details in data analysis and do not affect the qualitative conclusions we draw. Additionally, we performed non-preregistered validation analyses that were suggested by the main results.

2.1 Sample and session selection

The present analysis aimed to predict future cognitive decline from baseline non-brain (e.g., age and clinical scores) and brain imaging data (regional brain volume and functional connectivity). We used data from the publicly available, longitudinal OASIS-3 project, a collection of data from several studies at the Washington University Knight Alzheimer Disease Research Center (LaMontagne et al. 2019). OASIS-3 acquired data in different types of sessions (*clinical sessions:* non-brain data describing personal characteristics, cognitive and everyday functioning, health; *neuropsychological sessions:* non-brain data

from neuropsychological tests; *MRI sessions:* structural and functional MRI). The count and spacing between sessions varied between participants. To predict future cognitive decline, baseline sessions were used as input data and follow-up sessions as targets. The study design required a matching approach to select: i) baseline sessions (from clinical, neuropsychological, and MRI sessions) to be used as input data, and ii) follow-up clinical sessions to estimate the future cognitive decline.

First, baseline data were established by matching sessions from the different types (clinical, neuropsychological, MRI). We matched each MRI session that had at least one T1w and one fMRI scan with the closest clinical session. For each participant, the first MRI-clinical-session pair with an absolute time difference < 1 year was selected as baseline session. If no such pair was available, the participant was excluded from the analysis. Additionally, the closest neuropsychological session (within 1 year of the MRI baseline session) was also considered as baseline data. Baseline information from neuropsychological testing, however, was considered optional and not finding a matching neuropsychological session was not a criterion for exclusion. All data preceding the selected baseline sessions were disregarded for the analysis.

Second, all clinical sessions after the baseline clinical session were included as follow-up sessions to estimate cognitive decline. To reliably estimate decline, participants were only included if they had at least three clinical sessions (baseline plus two follow-up sessions). This matching approach reduced the sample (N_{total}=1098) to 662 participants (302 male; Table 2)[1]. The majority was cognitively healthy at baseline (509 healthy controls, 12 were diagnosed with MCI, and 111 with dementia; for 30 no diagnosis was available for the baseline session).

[1] The selected participants and session can be found here:

https://github.com/fliem/cpr/tree/0.1.2/info

MRI data was downloaded in BIDS format (K. J. Gorgolewski et al. 2016) via scripts provided by the OASIS project[2]. Non-brain data was downloaded via XNAT central[3].

[2] https://github.com/NrgXnat/oasis-scripts

[3] https://central.xnat.org/

Table 2. Sample characteristics. N = 662 (302 male). See Table 3 for a list of abbreviations.

	M	SD	min	max	N missing
Demographics					
$Age_{baseline}$	71	8.2	46	96	
Sex		302 M; 360 F			
Years of $education_{baseline}$	15	2,7	7	29	11
Clinical scores					
$MMSE_{baseline}$	28	2.2	16	30	
$CDR\text{-}SOB_{baseline}$	0.62	1.37	0.00	8.00	
$FAQ_{baseline}$	1.35	3.38	0.00	23.0	14
NPI-Q					
$Presence_{baseline}$	0.95	1.64	0.00	10.0	15
$Severity_{baseline}$	1.4	2.8	0.0	18	15
$GDS_{baseline}$	1.5	2.0	0.0	12	17
Genotyping (APOE alleles)					
E2 count		Zero: 565; One: 91; Two: 5			1
E3 count		Zero: 61; One: 272; Two: 328			1
E4 count		Zero: 403; One: 223; Two: 35			1
Diagnostic and sessions					
Clinical $diagnosis_{baseline}$		509 HC; 12 MCI; 111 DE			30
N clinical sessions	5.8	2.4	3	15	
Years between clinical sessions	1.2	0.6	0.003	5.5	
Years in study	5.8	2.5	1.6	10.9	
Outcomes					
MMSE slope (1/year)	-0.31	0.93	-7.5	2.9	
CDR-SOB slope (1/year)	0.25	0.61	-1.3	4,4	

2.2 Data

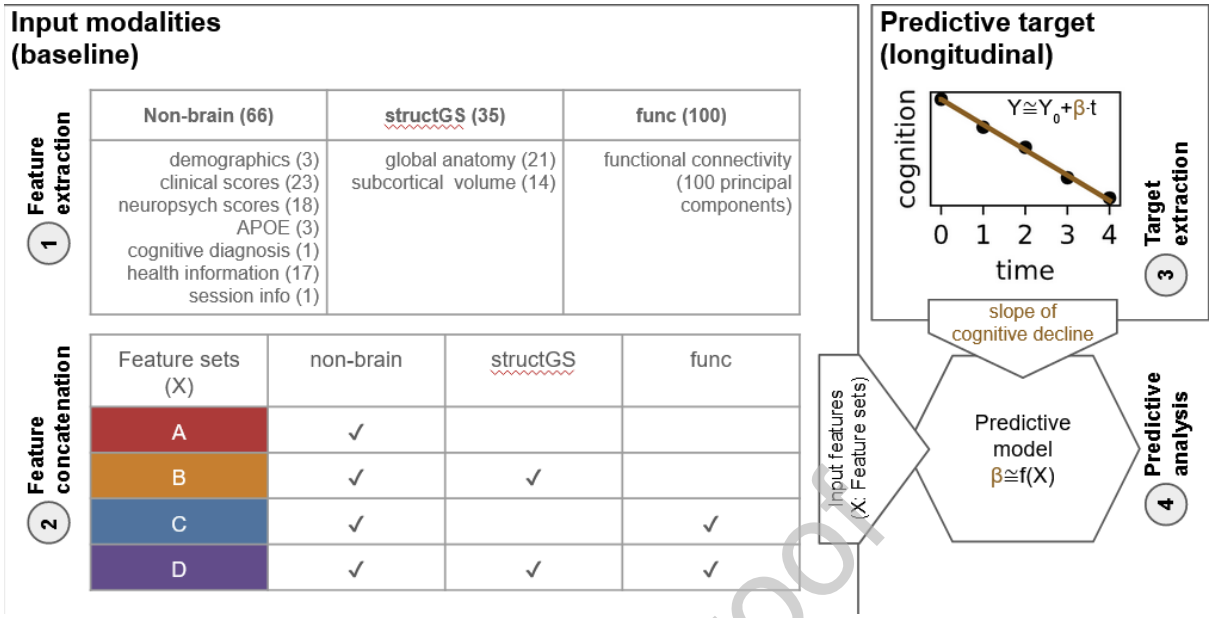

Figure 1. Overview of the predictive approach**.** 1) Features from *non-brain*, *structGS* (global and subcortical structural), and *func* (functional connectivity) modalities are extracted from baseline data. The number of features is provided in parentheses. 2) Feature concatenation produces sets of multimodal input features. For instance, red represents *non-brain* features only, while orange represents a combination of *non-brain* and *structGS*. 3) Extraction of slopes representing cognitive change from CDR (Clinical Dementia Rating) and MMSE (Mini-Mental State Examination). 4) Models are trained to predict cognitive decline based on the input features. Here, we used a multi-target random forest model within a nested cross-validation approach to predict CDR and MMSE change simultaneously.

2.2.1 Non-brain data

Non-brain data described personal characteristics at baseline, such as demographics, cognitive and everyday functioning, genetics, and health (Table 3 shows the abbreviations of tests). For further information on the measurements, see relevant publications by the OASIS team (LaMontagne et al. 2019; John C. Morris et al. 2006; Weintraub et al. 2009).

The specific measures included:

1. demographic information: sex, age, education
2. clinical scores: MMSE (Folstein, Folstein, and McHugh 1975), CDR (J. C. Morris 1993), FAQ (Jette et al. 1986), NPI-Q (Kaufer et al. 2000), GDS (Geriatric Depression Scale, Yesavage et al. 1982)
3. neuropsychological scores: WMS-R (Elwood 1991), Word fluency, TMT (Heller et al. 2013), WAIS-R (Franzen 2000), BNT (Borod, Goodglass, and Kaplan 1980)
4. APOE genotype
5. a cognitive diagnosis (healthy control, MCI, dementia)
6. health information: cardio/cerebrovascular health, diabetes, hypercholesterolemia, smoking, family history of dementia
7. the number of clinical sessions conducted before the selected baseline session (for instance sessions without a matching MRI session) to account for retest effects

Table 3. List of abbreviations of clinical tests

CDR-SOB	Clinical Dementia Rating - Sum of Boxes
FAQ	Functional Activities Questionnaire
GDS	Geriatric Depression Scale
MMSE	Mini-Mental State Examination
NPI-Q	Neuropsychiatric Inventory Questionnaire
TMT	Trail Making Test
WAIS-R	Wechsler Adult Intelligence Scale-Revised
WMS-R	Wechsler Memory Scale-Revised
BNT	Boston Naming Test

2.2.2 MRI data

MRI data were acquired on Siemens 3T scanners, with the majority coming from a TrioTim model (622 of 662 participants), and the rest from the combined PET/MRI Biograph mMR model. Each participant had between 1 and 4 T1w scans (1.7 on average). In total, the

sample had 1’119 T1w images. The parameter combination most used (in over 1’070 scans) was voxel size = 1 x 1 x 1 mm^3, echo time (TE) = 0.003 s, repetition time (TR) = 2.4 s. Where available, T2w images were also used to aid surface reconstruction. In total, 618 participants had a T2w image. The parameters for the T2w images were voxel size = 1 x 1 x 1 mm^3, TE = 0.455 s, TR = 3.2 s.

Each participant had between 1 and 4 functional resting-state scans (M = 2.0). In total, the sample had 1’327 functional images. The parameter combination most used (in over 1’300 scans) was voxel size = 4 x 4 x 4 mm^3, TE = 0.027 s, TR = 2.2 s, scan duration = 6 min. For further information regarding the imaging data see (LaMontagne et al. 2019).

See Table S1 for complete information about T1w, T2w and fMRI scans acquisitions.

2.3 MRI preprocessing

Functional and structural MRI data were preprocessed using the standard processing pipeline of *fMRIPrep* 1.4.1 (Esteban, Markiewicz, et al. 2018), which also includes running *FreeSurfer* 6.0.1 on the structural images (Fischl 2012). A detailed description of the preprocessing can be found in the supplement (*6.1.1 Details on MRI preprocessing*). Except for basic validity checks in a random subset of participants, data quality of the preprocessed data was not rigorously assessed. Notably *fMRIPrep* has been shown to robustly work across many datasets (Esteban, Markiewicz, et al. 2018).

2.4 Feature extraction

Input data from non-brain and brain imaging modalities at baseline were used to predict future cognitive decline (predictive targets). In the following sections we provide further details on the features that entered the predictive models.

2.4.1 Input data

Input data for the predictive models came from three modalities: *non-brain*, global and subcortical structural (*structGS*), and functional connectivity (*func*; Figure 1-1). Modalities

were entered into the models on their own and in combination. For instance, *non-brain* + *structGS* models received horizontally concatenated input features from the *non-brain* and *structGS* modalities (Figure 1-2). This allowed testing whether combining non-brain with structural data improved predictive accuracy as compared to non-brain data alone. The following paragraphs describe the input data modalities and Table S2 gives an overview of features entered into the models.

2.4.1.1 Non-brain data

Non-brain features included demographics, scores of clinical and neuropsychological instruments, APOE genotype, and health information. For a detailed list see Table S2. In total, 66 features entered the models from the *non-brain* modality.

2.4.1.2 Structural MRI (structGS)

For the *structGS* modality (global and subcortical structure), anatomical markers were extracted from the *FreeSurfer*-preprocessed anatomical scans. Following our previous work (Liem et al. 2017), we extracted global structural markers (volume of cerebellar and cerebral GM and WM, subcortical GM, ventricles, corpus callosum, and mean cortical thickness) and the volumes of seven subcortical regions (accumbens, amygdala, caudate, hippocampus, pallidum, putamen, thalamus; for each hemisphere separately). Most markers were extracted from the *aseg* file, except for mean cortical thickness, which was extracted from the *aparc.a2009s* parcellation (Desikan et al. 2006). To account for head-size-effects, volumetric values were normalized by estimated total intracranial volume. In total, 35 features entered the models from the *structGS* modality.

2.4.1.3 Functional MRI (func)

Functional connectivity was computed from the *fMRIPrep*-preprocessed functional scans. Denoising was performed using the 36P model (Ciric et al. 2017), which includes signals from 6 motion parameters, global, white matter, and CSF signals, derivatives, quadratic

terms, and squared derivatives. Time series were extracted from 300 cortical, cerebellar, and subcortical coordinates of the Seitzman atlas (Seitzman et al. 2020) using balls of 5 mm radius. The signals were band-pass filtered (0.01-0.1 Hz) and linearly detrended. Connectivity matrices were extracted by correlating the time series using Pearson correlation and applying Fisher-z-transformation. If multiple fMRI runs were available, the z-transformed connectivity matrices were averaged within participants. The vectorized upper triangle of this connectivity matrix was entered into the predictive pipeline and was further downsampled to 100 PCA components within cross-validation (see below). Denoising and feature extraction was performed with *Nilearn* 0.6.0 (Abraham et al. 2014).

Due to the dimensionality of connectivity matrices, we opted to perform PCA-based dimensionality reduction on the fMRI data. In its raw form, the fMRI connectivity has 44850 unique features, a prohibitive amount that eclipses the other modalities. Because trees are grown sampling both data points and variables, this number would complicate the training of Random Forests, requiring an increase in the number of trees to reliably expose the different modalities to the model. This problem is alleviated by reducing this feature set to 100 components. Because the degrees of freedom are greatly decreased, it also reduces the risk of overfitting. This solution has been successfully applied in many examples in the literature. Schulz et al. (2022), for example, uses PCA so that comparisons between modalities and modality combinations use the same number of features.

2.4.2 Predictive targets

To quantify future cognitive decline, trajectories of two clinical assessments, the CDR (Clinical Dementia Rating, Sum of Boxes score) and the MMSE (sum score of the Mini-Mental State Examination) were estimated using an ordinary least squares linear regression model for each participant and assessment independently (Figure 1-3; for information on the count and timing of sessions, see Table 2). A linear slope was fitted through the raw scores of the follow-up session with the intercept fixed at the raw score of the baseline session ($score_{follow-up,assessment} \sim score_{baseline,assessment} + \beta_{slope,assessment} \times time$). This approach

was chosen over a linear mixed effects model, as the mixed effects model requires data from multiple participants, making cross-validation more convoluted. The resulting two parameters ($\beta_{slope,CDR}$ and $\beta_{slope,MMSE}$) were the two targets that were simultaneously predicted in the predictive analysis using a multi-target approach (Rahim et al. 2017). Slopes were estimated with *Statsmodels* 0.10.1 (Seabold and Perktold 2010). The distribution of the estimated targets is plotted in Figure S1.

Two factors were fundamental to the adoption of the CDR and MMSE slopes as predictive targets. First, the number of participants with CDR and MMSE baseline scores is slightly higher than the number of participants with FAQ and NPI-Q scores. See Table 2. Second, both the CDR and MMSE are widely regarded as reliable tests for the clinical assessment and staging of dementia. FAQ, for example, is a questionnaire designed for bedside assessment and research based on instrumental activities of daily life (Pfeffer et al. 1982), which entails a degree of subjectivity, and NPI-Q is a brief informant-based questionnaire for general neuropsychiatric assessment (Kaufer et al. 2000). We believe the CDR and MMSE were the best candidates as specific measures of cognitive status for these two reasons.

2.5 Predictive analysis

The predictive pipeline (Figure 1-4) consisted of a multivariate imputer (*Scikit-learn's* IterativeImputer) (Buck 1960) and a multi-target random forest (RF) regression model (Breiman 2001). Multivariate imputation has recently been shown to robustly work in combination with predictive models under different missingness scenarios (Josse et al. 2019). It comprises using all other input variables to estimate missing values in each input variable vector. The procedure is then repeated now including the previously imputed values as inputs for a number of iterations. RF is a non-parametric machine-learning algorithm based on ensembles of decision trees. Trees are trained in parallel over bootstrap samples,

also including the sampling of input variables. The final output is obtained by aggregating the outputs of a predetermined number of trees.

Predictive models were trained using nested cross-validation via a stratified shuffle-split (1000 splits, 80% training, 20% test participants, stratified by the targets). In the inner loop, the RF's hyperparameters were tuned via grid search on the training participants (the tree depth was selected among [3, 5, 7, 10, 15, 20, 40, 50, None], where None leads to fully grown trees; the criterion to measure the quality of an RF-split was tuned with 'mean squared error' and 'mean absolute error'). The best estimator was carried forward to determine its out-of-sample performance on the test participants. To derive an estimate of chance performance, null-models were also trained and evaluated with permuted target values. For each cross-validation split, we calculated the coefficient of determination (R^2) on the test predictions. All predictive analyses were performed using *Scikit-learn* 0.22.1 (Pedregosa et al. 2011).

Model comparison was used to determine whether one model offered better prediction accuracy than another (for instance, to check whether a given model outperformed the null model, or whether a model with added brain imaging data improved accuracy as compared to a model using only non-brain data). Model comparison in a cross-validation needs to take the dependence between splits into account, complicating statistical tests (Bengio and Grandvalet 2004). Thus, instead of calculating a formal statistical test, we calculated the number of splits for which the model in question outperformed the reference model, resulting in a percent value, with numbers close to 100% denoting models which robustly outperformed the reference model (Engemann et al. 2020).

To inspect which features contributed to a prediction, permutation importance was calculated (Breiman 2001). Permutation importance evaluates the effects of features on the predictive performance by permuting feature values. If shuffling a feature does decrease performance, it is considered important for the model. It must be noted that this approach might underestimate the importance of correlated features. This is attributable to the fact that when one of the features is permuted, the second one retains some of the information that

both shared. For example, this might happen with baseline MMSE and CDR-SOB scores, which are correlated in the general population (Balsis et al. 2015), and are both used in the non-brain data. Furthermore, learning curves were estimated to assess whether the number of participants in the analysis was sufficient. For these comparisons, models were trained with increasing sample size while observing the test performance.

We performed additional analyses to diagnose the predictive pipeline and present our results in context. First, to validate the pipeline and analysis code, the same predictive methodology was used to predict age, a strong and well-established effect (Liem et al. 2017). For this validation, age was removed from the input data and the approach followed in the main analysis was repeated using a ridge regression model and 200 cross-validation splits. All features were normalized for this analysis, since ridge regression is sensitive to the variance of individual features. Ridge regression was selected for this analysis because research shows that linear models tend to perform on par with non-linear models in age prediction (Schulz et al. 2020, 2022) and other benchmarks (Dadi et al. 2019), especially at the current sample size.

Second, to better compare our results with previous work that predicted decline using class labels, we repeated the original pipeline to classify extreme groups of participants that are cognitively stable vs. participants with cognitive decline using random forest classifiers and 200 cross-validation splits. Participants with CDR-SOB slopes > 0.25/year were labeled as declining (N = 156), and a randomly drawn equal number of participants without change in CDR-SOB (N = 366) were labeled as stable. The threshold of 0.25/year was selected based on statistical considerations: it is the mean of CDR-SOB slopes, as shown in Table 2 and it also corresponds to the 80th percentile, so that choosing a higher threshold would result in a much smaller number of cases available.

2.6 Open science statement

All data used in the analysis are publicly available via the OASIS-3 project (LaMontagne et al. 2019). The analysis plan was preregistered (Liem et al. 2019). All preprocessing and analyses were performed in *Python* using open-source software and the code for preprocessing and predictive analysis is publicly available (Liem 2020)[4]. Furthermore, a docker container which includes all software and code to reproduce the preprocessing and predictive analysis is also provided[5].

3 Results

3.1 Predicting cognitive decline

A combination of non-brain and structural data gave the best predictions of future cognitive decline. Adding structural data improved the prediction for both the CDR (median test performance R^2 increased from 0.36 to 0.42; Figure 2, red vs. orange; for a scatter plot showing true vs predicted values, see Figure S3) and the MMSE (0.32 to 0.34), as compared to predictions from non-brain data alone. This increase occurred in a large majority of splits (91% of splits for CDR, 78% for MMSE; Table S3). In contrast, adding functional connectivity features to non-brain features, or to *non-brain + structGS* features, slightly decreased predictive performance.

To tune the RF models to the given problem, hyperparameters were optimized in a grid search approach. Tuning curves showed the results to be robust across a wide range of hyperparameter settings (Figure S4). Furthermore, learning curves demonstrated a sufficient sample size in the current setting (Figure S5).

The models consistently outperformed null models. Comparing the predictions against a null model with permuted predictions showed that most modalities outperformed chance-

[4] http://github.com/fliem/cpr

[5] https://hub.docker.com/r/fliem/cpr

level in 100% of splits (Table S4). The predictions based on functional connectivity were an exception and outperformed null-models to a lesser degree (91% of splits for CDR, 73% for MMSE).

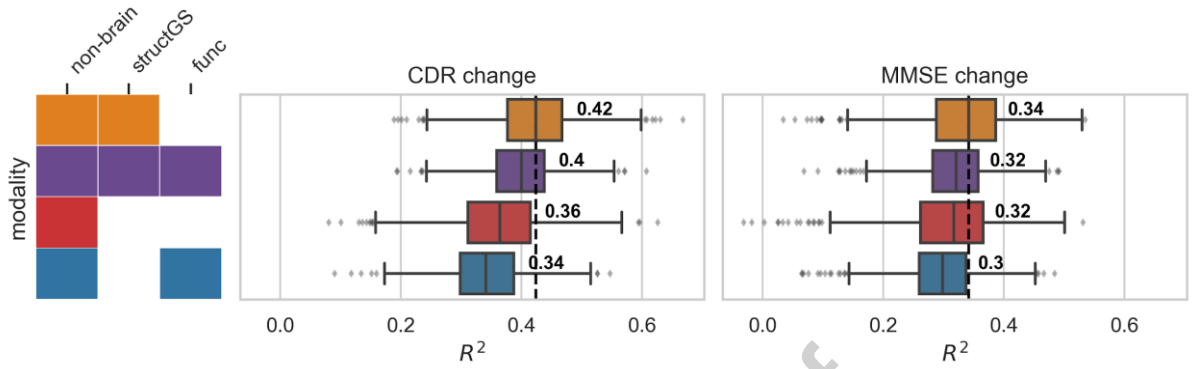

Figure 2**. Adding structural data (orange) to non-brain data (red) improved the prediction of cognitive decline.** Test performance (R^2, coefficient of determination, x-axis) across splits (N_{splits} = 1000) for the combinations of input modalities (y-axis). Targets: cognitive change measured via CDR (Clinical Dementia Rating, middle) and MMSE (Mini-Mental State Examination, right). Input modalities: non-brain, structGS (global and subcortical structural volumes), func (functional connectivity). The left panel represents combinations of input modalities (e.g., orange is *non-brain* + *structGS*). The number represents the median, the dashed vertical line marks the median of the best-performing combination of modalities (within a target measure). For the full results that include single-modality brain imaging, see Figure S2.

3.2 Features that predict cognitive decline

We used permutation importance to characterize the most predictive features of the best performing modality (non-brain + structGS). Within the top-15 features, non-brain included memory scores, the baseline scores of the targets (CDR, MMSE), and scores from the FAQ (functional assessment questionnaire). The structural features predominantly included subcortical regions (left and right hippocampus and amygdala, left accumbens; Figure 3).

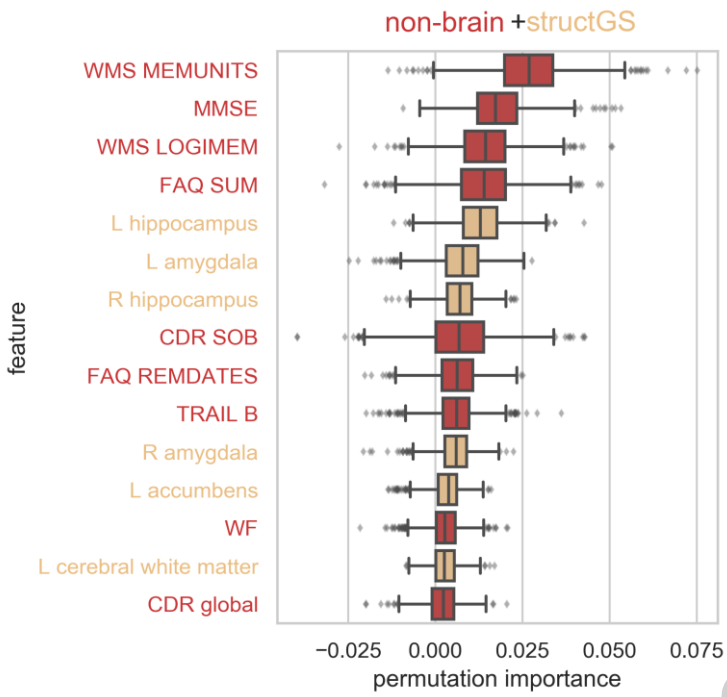

Figure 3. **Cognitive performance, daily functioning, and subcortical volume were among the most informative features.** Permutation importance of the top 15 features of the *non-brain + structGS* model (median across splits). Permutation importance is quantified as the decrease in test performance R^2 with the feature permuted. Red: non-brain features, light orange: structGS features. CDR: Clinical Dementia Rating, SOB: Sum of Boxes, FAQ: Functional Assessment Questionnaire, REMDATES: difficulty remembering dates, L: left, MMSE: Mini-Mental State Examination, R: right, SOB: sum of boxes score, TRAIL B: Trail Making Test B, WF: word fluency, WMS: Wechsler Memory Scale, MEMUNITS: Total number of story units recalled (delayed), LOGIMEM: Total number of story units recalled from this current test administration.

3.3 Validation analyses

Although functional connectivity models predicted cognitive decline poorly, functional data improved accuracy when predicting brain-age. Since functional connectivity alone did not predict cognitive decline well and did not increase the predictive accuracy of the non-brain model (Figure S2), we conducted a validation analysis to ensure that our functional connectivity models were able to predict brain-age, a well-established surrogate biomarker. Here, we predicted age from the same input data as in the main analysis after first removing

age from the input features set. In line with our expectations, functional connectivity increased predictive performance when combined with other modalities (e.g., in combination with non-brain, performance increased from 0.45 to 0.53; Figure 4), and functional connectivity alone could predict age reasonably well (median R^2 = 0.33, Figure S6), suggesting that its negligible contribution to decline prediction cannot be attributed to general methodological or data quality issues.

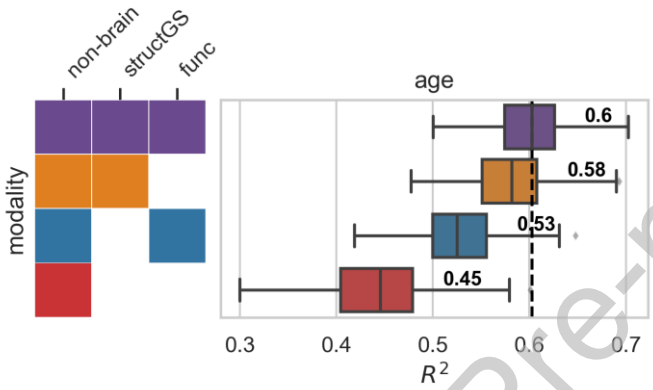

Figure 4. **Multimodal imaging improves brain-age prediction.** Input modalities: non-brain, structGS (global and subcortical structural volumes), func (functional connectivity). The number represents the median, the dashed vertical line marks the median of the best-performing combination of modalities. For the full results that include single-modality brain imaging, see Figure S6.

In the main analysis, the predictive target of cognitive decline was quantified as a continuous score. To compare our analysis to previous work that predicted classes of cognitive decline, we performed a further analysis that predicted extreme groups of cognitive decline (stable vs decline). Overall, extreme groups could be accurately predicted from the input data (most F1-scores [harmonic mean of the precision and recall] in the range of 0.8-0.9; Figure S7).

4 Discussion

In the present study, we found that combining baseline structural brain imaging data with non-brain data improved the prediction of future cognitive decline. In contrast, functional connectivity features did not improve prediction. By predicting future cognitive decline as a continuous trajectory, rather than a diagnostic label, our study broadens the scope of applications to cognitive decline in healthy aging. It also allows for more nuanced predictions on an individual level. In the future, these continuous measures may facilitate dimensional approaches to pathology (Cuthbert 2014).

The benefit of combining structural with non-brain data found in the present study is well in line with previous work that predicted conversion from MCI to AD (Korolev et al. 2016), and classes of cognitive decline (Bhagwat et al. 2018). Non-brain data alone predicted cognitive decline and the model was robustly improved by adding structural data (R^2 increased from 0.36 to 0.42 for CDR and from 0.32 to 0.34 for MMSE). These findings are consistent with prior work (Korolev et al. 2016; Bhagwat et al. 2018). In general, the range of accuracies reported in our study is well in line with previous work predicting a related continuous target (time to symptom onset in AD) (Vogel et al. 2018), as well as with work predicting diagnostic labels (e.g., Eskildsen et al. 2015; Korolev et al. 2016; Gaser et al. 2013; Davatzikos et al. 2011). After having established that a combination of non-brain and structural data gives predictions worthy of consideration, next, we assessed which features drove the predictions.

We found that clinical and neuropsychological assessments and subcortical structures drove the performance of our model. Measurements of memory, verbal fluency, executive function, and a wide set of cognitive and daily functions (MMSE, CDR, FAQ) were the most informative non-brain features for predicting cognitive decline. This matches well with Korolev et al. (2016) who found memory scores and clinical assessments (ADAS-Cog, FAQ) to be among the most informative non-brain features. On the other hand, hippocampus and amygdala volume were the most informative structural features in our analysis, which is well

in line with previous work predicting conversion from MCI to AD (Korolev et al. 2016; Eskildsen et al. 2015). In contrast, risk factors (such as age, APOE, or health risks) and markers that quantify general brain atrophy and regional cortical brain structure did not add markedly to model performance. It should be noted that features were assessed using permutation importance, which underestimates the importance of correlated features. Baseline MMSE and CDR-SOB scores are substantially correlated in the general population (Balsis et al. 2015), and are thus susceptible to this attenuation. We note, however, that both still figure among the top 15 features in Figure 3. Alternative approaches, such as mean decrease impurity, might complement the permutation-based approach in future studies to improve the sensitivity (Engemann et al. 2020). Nevertheless, taken together, our results suggest that memory, everyday functioning, and subcortical features better predict future cognitive decline at the individual level than risk factors or global brain characteristics.

Functional connectivity, in contrast to brain structure, did not improve predictions when added to other modalities, nor did it predict cognitive decline on its own. While many previous studies predicted cognitive performance or decline based on structural imaging, studies using functional connectivity are rare and contain widely varying estimates of its predictive power (Dansereau et al. 2017; Hojjati et al. 2018; Vogel et al. 2018). Although functional connectivity in our study did not predict future cognitive decline, it did predict brain-age. Assuming that functional connectivity is at least somewhat predictive of future cognitive decline, our analysis may suggest that the processing of functional connectivity data was not a good fit for the cognitive targets. Furthermore, data with better spatial and temporal resolution might be able to better capture decline. This calls for future studies that benchmark different processing options as these can severely impact predictive accuracy (Dubois et al. 2018).

In the following, we will sketch possible future developments along four themes: implications of and possible improvements to the continuous *targets of cognitive decline*, *multimodal input data*, *predictive models*, and the importance of *generalization* to new datasets.

Quantifying cognitive decline continuously rather than discretely enables a more fine-grained and robust prediction, but also requires methodological choices. By predicting a diagnostic label, previous studies were often restricted to MCI patients and aimed to distinguish stable from converting patients. Considering decline as a continuum better characterizes the underlying change in abilities and allows for capturing changes that occur in healthy aging. Overcoming the scarcity of diagnosed conditions, this widens the scope of applications and has methodological advantages: the resulting increased sample size yields more robust models, which is critical to avoid optimistic bias in estimating prediction accuracy (Woo et al. 2017; Varoquaux 2018). Furthermore, our approach also does not require assigning a diagnostic label, which entails subjective clinical judgment and arbitrary cut-off values. Considering cognitive decline as a continuous target does, however, require a model to aggregate multiple longitudinal measurements. Here, we used participant-specific linear slopes estimated through longitudinal data from clinical assessments. Since cognitive decline also shows nonlinear trajectories (Wilkosz et al. 2010), one could argue that accounting for nonlinearity is called for when extracting the predictive targets. However, robustly estimating nonlinearity requires more longitudinal measurements per participant and more complex models. In contrast, linear trajectories can robustly be estimated with three measurements, hence, they provide a useful approximation of cognitive decline. Notably, the baseline values of the clinical assessments used to define the slopes have a special role: they are input features and the slopes are defined relative to them. This might result in a bias due to regression to the mean (Barnett, van der Pols, and Dobson 2005), where unusually extreme baseline values (due to noise) might result in unusually extreme slopes (returning to the mean). This issue is relevant as well when defining diagnostic labels where it might result in patients switching between labels due to noise. Future studies should consider more complex models that can better account for these effects. Taken together, quantifying cognitive decline continuously allows for a more nuanced representation of decline and widens the scope of applications. However, while refining the definition of cognitive decline is warranted, it requires more complex analytical approaches and appropriate data.

In this study, we quantified cognitive decline using two clinical assessments (CDR and MMSE), which measure a heterogeneous set of cognitive and everyday life functions. While these clinical assessments have the advantage of being used in practice, they lack the specificity to target single cognitive constructs. Measuring cognitive constructs more homogeneously might potentially improve accuracy, especially if those constructs are strongly linked to specific brain regions or networks. This could be achieved by additionally employing neuropsychological assessments. The multi-target approach outlined in this study is well-suited to including these additional targets.

Beside additional targets, future studies should also consider additional multimodal input data to characterize the brain in greater detail. The present study used data derived from structural and functional MRI (T1w and resting-state fMRI). These might be complemented by information from diffusion-weighted imaging, arterial spin labeling, or positron emission tomography (Rahim et al. 2016). Additionally, the presently used modalities could also be refined and alternative representations could be considered. For instance, different methods for quantifying brain structure (Pipitone et al. 2014) or brain function (Rahim, Thirion, and Varoquaux 2019), and adding data on structural asymmetry (Wachinger et al. 2016) or dynamic functional connectivity (Filippi et al. 2019) could provide improved predictive performance. Furthermore, the influence of MR data quality on accuracy should be assessed in future studies. While our past work showed that brain-age prediction from multimodal neuroimaging is robust against in-scanner head motion (Liem et al. 2017), the present study has not assessed the influence of MR data quality on predictive accuracy. Addressing this issue would yield recommendations regarding the required data quality to predict cognitive decline.

The predictive approach could also be expanded to better accommodate high-dimensional data and the messiness of real-world data acquisition. The present study concatenated low-dimensional features across modalities and fed them into one random forest model. Including all features in one model allowed us to consider feature-level interactions across modalities. Alternatively, *prediction stacking* could be used to facilitate

the integration of multimodal data (Liem et al. 2017; Rahim et al. 2016; Engemann et al. 2020). While the stacking approach accounts for modality-level interactions it does not consider feature-level interactions across modalities. However, it scales well to high-dimensional data and allows for block-wise missing data, for instance, a missing modality. The present work only included participants if data from all modalities (non-brain, structural, functional) were available. In clinical practice, this is often not feasible. As we demonstrated previously, stacking can be used to include participants with missing modalities, which increases the sample size and the scope of application (Engemann et al. 2020).

In practice, the benefit of adding multimodal neuroimaging data to a set of clinical assessments needs to be considered against the additional costs. Its clinical utility also depends on the actionable insight that can be drawn from an earlier prognosis. Of course, this concern is not specific to this study; it applies broadly to almost every effort to incrementally predict clinically meaningful outcomes from brain-based measures. At the moment, no causal treatment of cognitive decline is available. However, an early prognosis might aid intervention studies and be even more helpful once effective treatments are available. Hence, future studies should further exploit the information yielded by the model to focus on participant-specific predictions. In general, predictive models don't perform equally well in all circumstances. For some participants or sub-groups, a more confident prediction is possible. Recent work demonstrated a higher prediction accuracy in participants with certain characteristics, e.g., older, female, etc. (Korolev et al. 2016). This enables increased accuracy by focusing on high-confidence predictions (Tam et al. 2019) and might even suggest a participant-tailored clinical workflow depending on the prediction confidence (Bhagwat et al. 2018). While the present study has not yet investigated these effects, it is well set-up to determine optimal conditions for model performance. The large number of cross-validation splits yields a distribution of predictive performance, not only a point estimate. This will also allow us to assess whether the predictions across sub-groups are driven by the same features.

For a predictive model to be useful in real-world applications, it needs to generalize well to datasets from different sites (Scheinost et al. 2019). While characteristics of our study facilitate generalization, a future study is required to empirically establish the generalization of our models to independent datasets. First, we have aimed to provide full transparency throughout this study to improve reproducibility and generalizability. We used data from a large, publicly available dataset, preprocessed them with well-established open-source tools and inputted them into well-established models. The analysis code is publicly shared and after further developing this approach, trained models will also be shared. Importantly, the analysis was preregistered to avoid overfitting due to analytical flexibility (Carp 2012; Hosseini et al. 2020). Second, the OASIS-3 project is set up heterogeneously regarding the number of sessions, the intervals between sessions, and the participants' duration in the study. This heterogeneity is expected to provide less opportunity for an algorithm to overfit to dataset-specific idiosyncrasies, resulting in more generalizable models that also perform well in other settings.

While a heterogeneous dataset and open/reproducible approaches certainly improve generalizability, we trained and tested models using only one dataset. Thus, the cross-validated performance in our study provides a biased estimate of the generalizability to independent datasets. This bias might even be modality-specific, in that non-brain features might generalize better than brain imaging features (Bhagwat et al. 2018). Training predictive models on data from multiple sites has been shown to improve generalization (Abraham et al. 2017; Orban et al. 2018; Liem et al. 2017). Hence, future studies should use models trained and tested on data from multiple sites, which requires further suitable longitudinal and publicly available datasets (Varoquaux 2018). This also provides an opportunity to take preregistration even further. After conducting experiments in an initial dataset, a trained model could be preregistered and applied to an independent dataset that hasn't yet been analyzed.

Future research could investigate the impact of different preprocessing strategies on predictive performance. This encompasses everything from individual settings to the

software used itself. For example, on the volumetry of subcortical structures, substantial differences are noted between different software (Mulder et al. 2014; Bartel et al. 2017).

5 Conclusions

In summary, we have shown that adding structural brain imaging data to non-brain data (such as memory scores or everyday functioning) improves the prediction of future cognitive decline in healthy and pathological aging. Conversely, adding functional connectivity data, as used in the present approach, did not aid the prediction. Importantly, our work has potential for clinical utility by predicting *future* cognitive decline, rather than a *current* diagnosis. Future studies should include additional brain imaging modalities and independent datasets and should determine the potential of functional connectivity using alternative methodological approaches. Quantifying future decline continuously allows for more nuanced predictions on an individual level. In the future, these continuous measures may facilitate dimensional approaches to pathology (Cuthbert 2014).

Increased personal and societal costs due to healthy and pathological age-related cognitive decline are one of the most pressing challenges in an aging society. An early and individually fine-grained prognosis of age-related cognitive decline allows for earlier and individually targeted behavioral, cognitive, or pharmacological interventions. Intervening early increases the chances to attenuate or prevent cognitive decline, which will alleviate both personal and societal costs. Importantly, our work targets applications to healthy aging, widening the scope beyond the pathological to the entire aging population.

Credit authorship statement

Bruno Hebling Vieira: Conceptualization, Writing - Review & Editing, Data Curation, Visualization; Franziskus Liem: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing - Original Draft, Writing - Review & Editing; Kamalaker Dadi: Conceptualization, Methodology, Writing - Review & Editing; Denis A. Engemann: Conceptualization, Methodology, Writing - Review & Editing; Alexandre Gramfort: Conceptualization, Methodology, Writing - Review & Editing; Pierre Bellec: Conceptualization, Writing - Review & Editing; R. Cameron Craddock: Conceptualization, Writing - Review & Editing; Jessica S. Damoiseaux: Conceptualization, Writing - Review & Editing; Christopher J. Steele: Conceptualization, Writing - Review & Editing; Tal Yarkoni:

Conceptualization, Writing - Review & Editing; Nicolas Langer: Conceptualization, Writing - Review & Editing; Daniel S. Margulies: Conceptualization, Writing - Review & Editing; Gaël Varoquaux: Conceptualization, Methodology, Writing - Review & Editing;

Acknowledgments
This work was supported by the URPP “Dynamics of Healthy Aging” and by the Swiss National Science Foundation [10001C_197480] at the University of Zurich and partially supported by the grant ANR-20-IADJ-0002 AI-cog to Alexandre Gramfort. Data were provided by the OASIS-3 project (Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50AG00561, P30NS09857781, P01AG026276, P01AG003991, R01AG043434, UL1TR000448, R01EB009352). We thank the participants and organizers of the OASIS-3 project for providing the data.
Conflict of Interest
The authors declare no conflict of interest

Verification

The authors verify that the data contained in the manuscript being submitted have not been previously published (except as a preprint on biorxiv.org), have not been submitted elsewhere and will not be submitted elsewhere while under consideration at Neurobiology of Aging.

References

Abraham, A., M.P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux. 2017. “Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example.” *NeuroImage* 147 (February): 736–45. https://doi.org/10.1016/j.neuroimage.2016.10.045.

Abraham, A., F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. 2014. “Machine learning for neuroimaging with Scikit-Learn.” *Frontiers in Neuroinformatics* 8. https://doi.org/10.3389/fninf.2014.00014.

Avants, B. B., C. L. Epstein, M. Grossman, and J. C. Gee. 2008. “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain.” *Medical Image Analysis* 12 (1): 26–41. https://doi.org/10.1016/j.media.2007.06.004.

Balsis, S., J.F. Benge, D.A. Lowe, L. Geraci, and R.S. Doody. 2015. “How do scores on the ADAS-Cog, MMSE, and CDR-SOB correspond?” *The Clinical Neuropsychologist* 29 (7): 1002–9. https://doi.org/10.1080/13854046.2015.1119312.

Barnett, A.G., J.C. van der Pols, and A.J. Dobson. 2005. “Regression to the mean: what it is and how to deal with it.” *International Journal of Epidemiology* 34 (1): 215–20. https://doi.org/10.1093/ije/dyh299.

Bartel, F., H. Vrenken, F. Bijma, F. Barkhof, M. van Herk, and J.C. de Munck. 2017. “Regional analysis of volumes and reproducibilities of automatic and manual hippocampal segmentations.” *PloS One* 12 (2): e0166785. https://doi.org/10.1371/journal.pone.0166785.

Bengio, Y., and Y. Grandvalet. 2004. “No unbiased estimator of the variance of k-fold cross-validation.” *Journal of Machine Learning Research: JMLR* 5 (Sep): 1089–1105. http://www.jmlr.org/papers/v5/grandvalet04a.html?92f58540.

Bhagwat, N., J.D. Viviano, A.N. Voineskos, M. Mallar Chakravarty, and Alzheimer’s Disease Neuroimaging Initiative. 2018. “Modeling and prediction of clinical symptom trajectories in Alzheimer’s disease using longitudinal data.” *PLoS Computational Biology* 14 (9): e1006376. https://doi.org/10.1371/journal.pcbi.1006376.

Borod, J.C., H. Goodglass, and E. Kaplan. 1980. “Normative data on the Boston diagnostic aphasia examination, parietal lobe battery, and the Boston Naming Test.” *Journal of Clinical Neuropsychology*. https://doi.org/10.1080/01688638008403793.

Breiman, L.. 2001. “Random Forests.” *Machine Learning*. https://doi.org/10.1023/A:1010933404324.

Buck, S.F. 1960. “A method of estimation of missing values in multivariate data suitable for use with an electronic computer.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 22 (2): 302–6. http://www.jstor.org/stable/2984099.

Carp, J.. 2012. “On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments.” *Frontiers in Neuroscience* 6 (October): 149. https://doi.org/10.3389/fnins.2012.00149.

Ciric, R., D.H. Wolf, J.D. Power, D.R. Roalf, G.L. Baum, K. Ruparel, R.T. Shinohara, et al. 2017. “Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity.” *NeuroImage* 154 (July): 174–87. https://doi.org/10.1016/j.neuroimage.2017.03.020.

Cole, J.H., and K. Franke. 2017. “Predicting age using neuroimaging: innovative brain ageing biomarkers.” *Trends in Neurosciences* 40 (12): 681–90. https://doi.org/10.1016/j.tins.2017.10.001.

Cuthbert, B.N. 2014. “The RDoC Framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology.” *World Psychiatry: Official Journal of the World Psychiatric Association* 13 (1): 28–35. https://doi.org/10.1002/wps.20087.

Dadi, K., M. Rahim, A. Abraham, D. Chyzhyk, M. Milham, B. Thirion, G. Varoquaux, and Alzheimer’s Disease Neuroimaging Initiative. 2019. “Benchmarking functional connectome-based predictive models for resting-state fMRI.” *NeuroImage* 192 (May): 115–34. https://doi.org/10.1016/j.neuroimage.2019.02.062.

Dadi, K., G. Varoquaux, J. Houenou, D. Bzdok, B. Thirion, and D. Engemann. 2021. “Population modeling with machine learning can enhance measures of mental health.” *GigaScience* 10 (10). https://doi.org/10.1093/gigascience/giab071.

Dale, A.M., B. Fischl, and M.I. Sereno. 1999. “Cortical surface-based analysis: I. segmentation and surface reconstruction.” *NeuroImage* 9 (2): 179–94. https://doi.org/10.1006/nimg.1998.0395.

Dansereau, C., A. Tam, A. Badhwar, S. Urchs, P. Orban, P. Rosa-Neto, and P. Bellec. 2017. “A brain signature highly predictive of future progression to Alzheimer’s dementia.” http://arxiv.org/abs/1712.08058.

Davatzikos, C.. 2019. “Machine learning in neuroimaging: progress and challenges.” *NeuroImage* 197 (August): 652–56. https://doi.org/10.1016/j.neuroimage.2018.10.003.

Davatzikos, C., P. Bhatt, L.M. Shaw, K.N. Batmanghelich, and J.Q. Trojanowski. 2011. “Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification.” *Neurobiology of Aging* 32 (12): 2322.e19–27. https://doi.org/10.1016/j.neurobiolaging.2010.05.023.

Desikan, R.S., F. Ségonne, B. Fischl, B.T. Quinn, B.C. Dickerson, D. Blacker, R.L. Buckner, et al. 2006. “An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.” *NeuroImage* 31 (3): 968–80. https://doi.org/10.1016/j.neuroimage.2006.01.021.

Dubois, J., P. Galdi, Y. Han, L.K. Paul, and R. Adolphs. 2018. “Resting-state functional brain connectivity best predicts the personality dimension of openness to experience.” *Personality Neuroscience*. https://doi.org/10.1017/pen.2018.8.

Elwood, R. W. 1991. “The Wechsler Memory Scale-Revised: psychometric characteristics and clinical application.” *Neuropsychology Review* 2 (2): 179–201. https://doi.org/10.1007/bf01109053.

Engemann, D.A., O. Kozynets, D. Sabbagh, G. Lemaître, G. Varoquaux, F. Liem, and A. Gramfort. 2020. “Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers.” *eLife* 9 (May). https://doi.org/10.7554/eLife.54055.

Eskildsen, S.F., P. Coupé, V.S. Fonov, J.C. Pruessner, D. Louis Collins, and Alzheimer’s Disease Neuroimaging Initiative. 2015. “Structural imaging biomarkers of Alzheimer’s disease: predicting disease progression.” *Neurobiology of Aging* 36 Suppl 1 (January): S23–31. https://doi.org/10.1016/j.neurobiolaging.2014.04.034.

Esteban, O., R. Blair, C.J. Markiewicz, S.L. Berleant, C. Moodie, F. Ma, A. Ilkay Isik, et al. 2018. “fMRIPrep.” *Software: Practice & Experience*. https://doi.org/10.5281/zenodo.852659.

Esteban, O., C. Markiewicz, R.W. Blair, C. Moodie, A. Ilkay Isik, A. Erramuzpe Aliaga, J. Kent, et al. 2018. “fMRIPrep: a robust preprocessing pipeline for functional MRI.” *Nature Methods*. https://doi.org/10.1038/s41592-018-0235-4.

Filippi, M., E.G. Spinelli, C. Cividini, and F. Agosta. 2019. “Resting state dynamic functional connectivity in neurodegenerative conditions: a review of magnetic resonance imaging findings.” *Frontiers in Neuroscience* 13 (June): 657. https://doi.org/10.3389/fnins.2019.00657.

Fischl, B.. 2012. “FreeSurfer.” *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2012.01.021.

Folstein, M.F., S.E. Folstein, and P.R. McHugh. 1975. “‘Mini-Mental State’. a practical method for grading the cognitive state of patients for the clinician.” *Journal of Psychiatric Research* 12 (3): 189–98. https://doi.org/10.1016/0022-3956(75)90026-6.

Fonov, V.S., A.C. Evans, R.C. McKinstry, C.R. Almli, and D.L. Collins. 2009. “Unbiased nonlinear average age-appropriate brain templates from birth to adulthood.” *NeuroImage* 47, Supplement 1: S102. https://doi.org/10.1016/S1053-8119(09)70884-5.

Franzen, Michael D. 2000. “The Wechsler Adult Intelligence Scale-Revised and Wechsler Adult Intelligence Scale-III.” *Reliability and Validity in Neuropsychological Assessment*. https://doi.org/10.1007/978-1-4757-3224-5_6.

Gaser, C., K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, and Alzheimer’s Disease Neuroimaging Initiative. 2013. “BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer’s disease.” *PloS One* 8 (6): e67346. https://doi.org/10.1371/journal.pone.0067346.

Gorgolewski, K., C.D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M.L. Waskom, and S. Ghosh. 2011. “Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python.” *Frontiers in Neuroinformatics* 5: 13. https://doi.org/10.3389/fninf.2011.00013.

Gorgolewski, K.J., T. Auer, V.D. Calhoun, R.C. Craddock, S. Das, E.P. Duff, G. Flandin, et al. 2016. “The Brain Imaging Data Structure, a format for organizing and describing outputs of neuroimaging experiments.” *Scientific Data* 3 (June): 160044. https://doi.org/10.1038/sdata.2016.44.

Gorgolewski, K.J., O. Esteban, C.J. Markiewicz, E. Ziegler, D.G. Ellis, M.P. Notter, D. Jarecka, et al. 2018. “Nipype.” *Software: Practice & Experience*. https://doi.org/10.5281/zenodo.596855.

Greve, D.N., and B. Fischl. 2009. “Accurate and robust brain image alignment using boundary-based registration.” *NeuroImage* 48 (1): 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

Grüne, A.. 2020. “What is a good F1 score?.” Inside GetYourGuide. October 13, 2020. https://inside.getyourguide.com/blog/2020/9/30/what-makes-a-good-f1-score.

Heller, L.J., C.S. Skinner, A.J. Tomiyama, E.S. Epel, P.A. Hall, J. Allan, L. LaCaille, et al. 2013. “Trail-making test.” In *Encyclopedia of Behavioral Medicine*, edited by Marc D. Gellman and J. Rick Turner, 1986–87. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-1005-9_1538.

Hojjati, S.H., A. Ebrahimzadeh, A. Khazaee, A. Babajani-Feremi, and Alzheimer’s Disease Neuroimaging Initiative. 2018. “Predicting conversion from MCI to AD by integrating Rs-fMRI and structural MRI.” *Computers in Biology and Medicine* 102 (November): 30–39. https://doi.org/10.1016/j.compbiomed.2018.09.004.

Hosseini, M., M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, and B. Wyble. 2020. “I tried a bunch of things: the dangers of unexpected overfitting in classification of brain data.” *Neuroscience and Biobehavioral Reviews* 119 (December): 456–67. https://doi.org/10.1016/j.neubiorev.2020.09.036.

Jenkinson, M., P. Bannister, M. Brady, and S. Smith. 2002. “Improved optimization for the robust and accurate linear registration and motion correction of brain images.” *NeuroImage* 17 (2): 825–41. https://doi.org/10.1006/nimg.2002.1132.

Jette, A. M., A. R. Davies, P. D. Cleary, D. R. Calkins, L. V. Rubenstein, A. Fink, J. Kosecoff, R. T. Young, R. H. Brook, and T. L. Delbanco. 1986. “The functional status questionnaire: reliability

and validity when used in primary care." *Journal of General Internal Medicine* 1 (3): 143–49. https://doi.org/10.1007/BF02602324.

Josse, J., N. Prost, E. Scornet, and G. Varoquaux. 2019. "On the consistency of supervised learning with missing values." http://arxiv.org/abs/1902.06931.

Kaufer, D.I., J.L. Cummings, P. Ketchel, V. Smith, A. MacMillan, T. Shelley, O.L. Lopez, and S.T. DeKosky. 2000. "Validation of the NPI-Q, a brief clinical form of the neuropsychiatric inventory." *The Journal of Neuropsychiatry and Clinical Neurosciences* 12 (2): 233–39. https://doi.org/10.1176/jnp.12.2.233.

Klein, A., S.S. Ghosh, F.S. Bao, J. Giard, Y. Häme, E. Stavsky, N. Lee, et al. 2017. "Mindboggling morphometry of human brains." *PLoS Computational Biology* 13 (2): e1005350. https://doi.org/10.1371/journal.pcbi.1005350.

Korolev, I.O., L.L. Symonds, A.C. Bozoki, and Alzheimer's Disease Neuroimaging Initiative. 2016. "Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification." *PloS One* 11 (2): e0138866. https://doi.org/10.1371/journal.pone.0138866.

LaMontagne, P.J., T.L.S. Benzinger, J.C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, et al. 2019. "OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease." *Radiology and Imaging*. medRxiv. https://doi.org/10.1101/2019.12.13.19014902

Lanczos, C. 1964. "Evaluation of noisy data." *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* 1 (1): 76–85. https://doi.org/10.1137/0701007.

Liem, F. 2020. *Fliem/cpr 0.1.1*. https://doi.org/10.5281/zenodo.3726641.

Liem, F., P. Bellec, C. Craddock, K. Dadi, J.S. Damoiseaux, D.S. Margulies, C.J. Steele, G. Varoquaux, and T. Yarkoni. 2019. "Predicting future cognitive change from multiple data sources (pilot Study)." OSF. https://doi.org/10.17605/OSF.IO/GYNBJ.

Liem, F., L. Geerligs, J.S. Damoiseaux, and D.S. Margulies. 2020. "Functional connectivity in aging." https://doi.org/10.31234/osf.io/whsud.

Liem, F., G. Varoquaux, J. Kynast, F. Beyer, S. Kharabian Masouleh, J.M. Huntenburg, L. Lampe, et al. 2017. "Predicting brain-age from multimodal imaging data captures cognitive impairment." *NeuroImage* 148 (March): 179–88. https://doi.org/10.1016/j.neuroimage.2016.11.005.

Morris, J.C. 1993. “The Clinical Dementia Rating (CDR): current version and scoring rules.” *Neurology* 43 (11): 2412–14. https://doi.org/10.1212/wnl.43.11.2412-a.

Morris, J.C., S. Weintraub, H.C. Chui, J. Cummings, C. Decarli, S. Ferris, N.L. Foster, et al. 2006. “The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer disease centers.” *Alzheimer Disease and Associated Disorders* 20 (4): 210–16. https://doi.org/10.1097/01.wad.0000213865.09806.92.

Mulder, E.R., R.A. de Jong, D.L. Knol, R.A. van Schijndel, K.S. Cover, P.J. Visser, F. Barkhof, H. Vrenken, and Alzheimer’s Disease Neuroimaging Initiative. 2014. “Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST.” *NeuroImage* 92 (May): 169–81. https://doi.org/10.1016/j.neuroimage.2014.01.058.

Na, K.-S.. 2019. “Prediction of future cognitive impairment among the community elderly: a machine-learning based approach.” *Scientific Reports* 9 (1): 3335. https://doi.org/10.1038/s41598-019-39478-7.

Orban, P., C. Dansereau, L. Desbois, V. Mongeau-Pérusse, C.-É. Giguère, H. Nguyen, A. Mendrek, E. Stip, and P. Bellec. 2018. “Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity.” *Schizophrenia Research* 192 (February): 167–71. https://doi.org/10.1016/j.schres.2017.05.027.

Oschwald, J., S. Guye, F. Liem, P. Rast, S. Willis, C. Röcke, L. Jäncke, M. Martin, and S. Mérillat. 2019. “Brain structure and cognitive ability in healthy aging: a review on longitudinal correlated change.” *Reviews in the Neurosciences* 31 (1): 1–57. https://doi.org/10.1515/revneuro-2018-0096.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. “Scikit-Learn: machine learning in Python.” *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–30. http://www.jmlr.org/papers/v12/pedregosa11a.html.

Pfeffer, R.I., T.T. Kurosaki, C.H. Harrah Jr, J.M. Chance, and S. Filos. 1982. “Measurement of functional activities in older adults in the community.” *Journal of Gerontology* 37 (3): 323–29. https://doi.org/10.1093/geronj/37.3.323.

Pipitone, J., M.T.M. Park, J. Winterburn, T.A. Lett, J.P. Lerch, J.C. Pruessner, M. Lepage, A.N. Voineskos, M. Mallar Chakravarty, and Alzheimer’s Disease Neuroimaging Initiative. 2014.

"Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates." *NeuroImage* 101 (November): 494–512. https://doi.org/10.1016/j.neuroimage.2014.04.054.

Power, J.D., A. Mitra, T.O. Laumann, A.Z. Snyder, B.L. Schlaggar, and S.E. Petersen. 2014. "Methods to detect, characterize, and remove motion artifact in resting state fMRI." *NeuroImage* 84 (Supplement C): 320–41. https://doi.org/10.1016/j.neuroimage.2013.08.048.

Rahim, M., B. Thirion, D. Bzdok, I. Buvat, and G. Varoquaux. 2017. "Joint prediction of multiple scores captures better individual traits from brain images." *NeuroImage* 158 (September): 145–54. https://doi.org/10.1016/j.neuroimage.2017.06.072.

Rahim, M., B. Thirion, C. Comtat, G. Varoquaux, and Alzheimer's Disease Neuroimaging Initiative. 2016. "Transmodal learning of functional networks for Alzheimer's disease prediction." *IEEE Journal of Selected Topics in Signal Processing* 10 (7): 120–1213. https://doi.org/10.1109/JSTSP.2016.2600400.

Rahim, M., B. Thirion, and G. Varoquaux. 2019. "Population Shrinkage of Covariance (PoSCE) for better individual brain functional-connectivity estimation." *Medical Image Analysis* 54 (May): 138–48. https://doi.org/10.1016/j.media.2019.03.001.

Rathore, S., M. Habes, M. Aksam Iftikhar, A. Shacklett, and C. Davatzikos. 2017. "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages." *NeuroImage* 155 (July): 530–48. https://doi.org/10.1016/j.neuroimage.2017.03.057.

Reuter, M., H.D. Rosas, and B. Fischl. 2010. "Highly accurate inverse consistent registration: a robust approach." *NeuroImage* 53 (4): 1181–96. https://doi.org/10.1016/j.neuroimage.2010.07.020.

Scheinost, D., S. Noble, C. Horien, A.S. Greene, E. Mr Lake, M. Salehi, S. Gao, et al. 2019. "Ten simple rules for predictive modeling of individual differences in neuroimaging." *NeuroImage* 193 (June): 35–45. https://doi.org/10.1016/j.neuroimage.2019.02.057.

Schulz, M.-A., D. Bzdok, S. Haufe, J.-D. Haynes, and K. Ritter. 2022. "Performance reserves in brain-imaging-based phenotype prediction." *bioRxiv Preprint*. https://doi.org/10.1101/2022.02.23.481601.

Schulz, M.-A., B.T. Thomas Yeo, J.T. Vogelstein, J. Mourao-Miranada, J.N. Kather, K. Kording, B. Richards, and D. Bzdok. 2020. “Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets.” *Nature Communications* 11 (1): 4238. https://doi.org/10.1038/s41467-020-18037-z.

Seabold, S., and J. Perktold. 2010. “Statsmodels: econometric and statistical modeling with Python.” In *Proceedings of the 9th Python in Science Conference*, 57:61. Scipy. https://doi.org/10.25080/Majora-92bf1922-011.

Seitzman, B.A., C. Gratton, S. Marek, R.V. Raut, N.U.F. Dosenbach, B.L. Schlaggar, S.E. Petersen, and D.J. Greene. 2020. “A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum.” *NeuroImage* 206 (February). https://doi.org/10.1016/j.neuroimage.2019.116290.

Tam, A., C. Dansereau, Y. Iturria-Medina, S. Urchs, P. Orban, H. Sharmarke, J. Breitner, P. Bellec, and Alzheimer’s Disease Neuroimaging Initiative. 2019. “A highly predictive signature of cognition and brain atrophy for progression to Alzheimer’s dementia.” *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz055.

Tustison, N.J., B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, and J.C. Gee. 2010. “N4ITK: improved N3 bias correction.” *IEEE Transactions on Medical Imaging* 29 (6): 1310–20. https://doi.org/10.1109/TMI.2010.2046908.

Varoquaux, G. 2018. “Cross-validation failure: small sample sizes lead to large error bars.” *NeuroImage* 180 (Pt A): 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061.

Vogel, J.W., E. Vachon-Presseau, A. Pichet Binette, A. Tam, P. Orban, R. La Joie, M. Savard, et al. 2018. “Brain properties predict proximity to symptom onset in sporadic Alzheimer’s disease.” *Brain: A Journal of Neurology* 141 (6): 1871–83. https://doi.org/10.1093/brain/awy093.

Wachinger, C., D.H. Salat, M. Weiner, M. Reuter, and Alzheimer’s Disease Neuroimaging Initiative. 2016. “Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala.” *Brain: A Journal of Neurology* 139 (Pt 12): 3253–66. https://doi.org/10.1093/brain/aww243.

Weintraub, S., D. Salmon, N. Mercaldo, S. Ferris, N.R. Graff-Radford, H. Chui, J. Cummings, et al. 2009. “The Alzheimer’s Disease Centers' Uniform Data Set (UDS): the neuropsychologic test

battery." *Alzheimer Disease and Associated Disorders* 23 (2): 91–101. https://doi.org/10.1097/WAD.0b013e318191c7dd.

Wilkosz, P.A., H.J. Seltman, B. Devlin, E.A. Weamer, O.L. Lopez, S.T. DeKosky, and R.A. Sweet. 2010. "Trajectories of cognitive decline in Alzheimer's disease." *International Psychogeriatrics / IPA* 22 (2): 281–90. https://doi.org/10.1017/S1041610209991001.

Woo, C.-W., L.J. Chang, M.A. Lindquist, and T.D. Wager. 2017. "Building better biomarkers: brain models in translational neuroimaging." *Nature Neuroscience* 20 (3): 365–77. https://doi.org/10.1038/nn.4478.

Yesavage, J. A., T. L. Brink, T. L. Rose, O. Lum, V. Huang, M. Adey, and V. O. Leirer. 1982. "Development and validation of a geriatric depression screening scale: a preliminary report." *Journal of Psychiatric Research* 17 (1): 37–49. https://www.ncbi.nlm.nih.gov/pubmed/7183759.

Zhang, Y., M. Brady, and S. Smith. 2001. "Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm." *IEEE Transactions on Medical Imaging* 20 (1): 45–57. https://doi.org/10.1109/42.906424.