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

Franziskus Liem1, Kamalaker Dadi2, Denis A. Engemann2,3, Alexandre Gramfort2, Pierre Bellec4, R. Cameron Craddock5, Jessica S. Damoiseaux6, Christopher J. Steele7, Tal Yarkoni5, Daniel S. Margulies8, Gaël Varoquaux2

1University Research Priority Program “Dynamics of Healthy Aging”, University of Zurich, Zurich, Switzerland
2Université Paris-Saclay, Inria, CEA, Palaiseau, France
3Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
4University of Montreal, Montreal, Canada
5The University of Texas, Austin, TX, USA
6Wayne State University, Detroit, MI, USA
7Concordia University, Montreal, Canada
8Institut du Cerveau et de la Moelle épinière, Paris, France

Cognitive decline occurs in healthy and pathological aging, and both may be preceded by subtle changes in the brain — offering a basis for cognitive predictions. Previous work has largely focused on predicting a diagnostic label from structural brain imaging. Our study broadens the scope of applications to cognitive decline in healthy aging by predicting future decline as a continuous trajectory, rather than a diagnostic label. Furthermore, since brain structure as well as function changes in aging, it is reasonable to expect predictive gains when using multiple brain imaging modalities. Here, we tested whether baseline multimodal neuroimaging data improve the prediction of future cognitive decline in healthy and pathological aging. Non-brain data (including demographics and clinical and neuropsychological scores) were combined with structural and functional connectivity MRI data from the OASIS-3 project (N = 662; age = 46 – 96y). The combined input data was entered into cross-validated multi-target random forest models to predict future cognitive decline (measured by the Clinical Dementia Rating and the Mini-Mental State Examination), on average 5.8y into the future. The analysis was preregistered and all analysis code is publicly available. We found that combining non-brain with structural data improved the continuous prediction of future cognitive decline (best test-set performance: $R^2 = 0.42$) and that cognitive performance, daily functioning, and subcortical volume drove the performance of our model. In contrast, including functional connectivity did not improve predictive accuracy. In the future, the prognosis of age-related cognitive decline may enable earlier and more effective cognitive, pharmacological, and behavioral interventions to be tailored to the individual.

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

Acknowledgments: This work was supported by the URPP “Dynamics of Healthy Aging” at the University of Zurich. 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.
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) (Davatzikos et al., 2011; e.g., Eskildsen et al., 2015; Gaser et al., 2013; Korolev et al., 2016) or predicted membership in data-driven trajectory-groups of future decline (Bhagwat et al., 2018). 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.

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 (Engemann et al., 2020; Liem et al., 2017). 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 (Dansereau et al., 2017; e.g., Hojjati et al., 2018). Therefore, integrating multiple brain imaging modalities enables a more complete characterization of brain aging and provides increased predictive power.

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 subjects. 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 subject, 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 subject 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, subjects were only included if they had at least three clinical sessions (baseline plus two follow-up sessions).

This matching approach reduced the sample \(N_{\text{total}}=1098\) to 662 subjects (302 male; Table 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).

MRI data was downloaded in BIDS format (Gorgolewski et al., 2016) via scripts provided by the OASIS project. Non-brain data was downloaded via XNAT central.

### Table 1. Sample characteristics. \(N=662\) (302 male). MMSE: Mini-Mental State Examination

|                      | M    | SD   | min | max |
|----------------------|------|------|-----|-----|
| age\(_{\text{baseline}}\) | 71   | 8    | 46  | 96  |
| MMSE\(_{\text{baseline}}\)  | 28   | 2    | 16  | 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 |

### 2.2 Data

#### 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 2 shows the abbreviations of tests). For further information on the measurements, see relevant publications by the OASIS team (LaMontagne et al., 2019; Morris et al., 2006; Weintraub et al., 2009).

The specific measures included:

1. demographic information: sex, age, education
2. clinical scores: MMSE (Folstein et al., 1975), CDR (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 et al., 1980)
4. APOE genotype
5. a cognitive diagnosis (healthy control, MCI, dementia)
6. health information: cardio/cerebro-vascular health, diabetes, hypercholesterolemia, smoking, family history of dementia
7. the number of clinical session conducted before the selected baseline session (for instance sessions without a matching MRI session) to account for retest effects

| CDR  | Clinical Dementia Rating |
| 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 |

Table 2. List of abbreviations of clinical tests

#### 2.2.2 MRI data

MRI data were acquired on Siemens 3T scanners, with the majority coming from a TrioTrim model (622 of 662 subjects). 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 commonly 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 commonly 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).

Input modalities (baseline)

| Feature extraction          | non-brain | structGS | func    |
|----------------------------|-----------|----------|---------|
| demographics               |           |          |         |
| clinical scores            |           |          |         |
| neuropsych scores          |           |          |         |
| APOE                       |           |          |         |
| cognitive diagnosis        |           |          |         |
| health                     |           |          |         |
| global anatomy             |           |          |         |
| subcortical volume         |           |          |         |
| functional connectivity    |           |          |         |

Predictive target (longitudinal)

| Target extraction          | Input features | Predictive model | Predictive analysis |
|----------------------------|----------------|------------------|--------------------|
| slope of cognitive decline  |                |                  |                    |

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. 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 the 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.

2.3 MRI preprocessing

Functional and structural MRI data were preprocessed using the standard processing pipeline of fMRIPrep 1.4.1 (Esteban et al., 2018b), 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 subjects, data quality of the preprocessed data was not rigorously assessed. Notably fMRIPrep has been shown to robustly work across many datasets (Esteban et al., 2018b).

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 S1 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 S1. 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 (Circi 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., 2018) 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 (0.01-0.1 Hz) and linearly detrended. Connectivity derivatives, quadratic terms, and squared derivatives. Parameters, global, white matter, and CSF signals, -preprocessed functional scans. Denoising fMRIPrep 2.4.1.3 Functional MRI (func)

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).

Predictive models were trained using nested cross-validation via a stratified shuffle-split (1000 splits, 80% training, 20% test subjects, stratified by the targets). In the inner loop, the RF’s hyperparameters were tuned via grid search on the training subjects (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 subjects. To derive an estimate of chance performance, null-models were also trained and tested with permuted target values. For each cross-validation split, the coefficient of determination (R²) was calculated 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 permuted 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 permuted model.
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 has to be noted that this approach might underestimate the importance of correlated features. Furthermore, learning curves were estimated to assess whether the number of subjects 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.

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 subjects that are cognitively stable vs. subjects with cognitive decline using random forest classifiers and 200 cross-validation splits. Subjects with CDR-SOB slopes > 0.25 were labeled as declining (N = 156), and a randomly drawn equal number of subjects without change in CDR-SOB were labeled as stable.

### 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\(^5\).

---

\(^4\) http://github.com/fliem/cpr

\(^5\) https://hub.docker.com/r/fliem/cpr

---

### 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 S2). 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-level in 100% of splits (Table S3). 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).

#### 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 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.

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, FAQ: Functional Assessment Questionnaire, 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.
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.

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).

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.

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 (Bhagwat et al., 2018; Korolev et al., 2016). 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 (Davatzikos et al., 2011; e.g., Eskildsen et al., 2015; Gaser et al., 2013; Korolev et al., 2016). 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 (Eskildsen et al., 2015; Korolev et al., 2016). 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. 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; Hoijati 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 (Varoquaux, 2018; Woo et al., 2017). 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 subject-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 subject 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 et al., 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 et al., 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 (Engemann et al., 2020; Liem et al., 2017; Rahim et al., 2016). 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 subjects 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 subjects 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 subject-specific predictions. In general, predictive models don’t perform equally well in all circumstances. For some subjects or sub-groups a more confident prediction is possible. Recent work demonstrated a higher prediction accuracy in subjects 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 subject-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; Skocik et al., 2016). 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; Liem et al., 2017; Orban et al., 2018). 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.

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.
6 Appendix

6.1 Supplementary Methods

6.1.1 Details on MRI preprocessing

Results included in this manuscript come from data preprocessed using fMRIPrep 1.4.1 (RRID:SCR_016216; Esteban et al., 2018a, 2018b), which is based on Nipype 1.2.0 (RRID:SCR_002502; Gorgolewski et al., 2011, 2018).

6.1.1.1 Anatomical data preprocessing

T1-weighted (T1w) images were corrected for intensity non-uniformity (INU) with ‘N4BiasFieldCorrection’ (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008; RRID:SCR_004757; Tustison et al., 2010). The T1w-reference was then skull-stripped with a Nipype implementation of the ‘antsBrainExtraction.sh’ workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using ‘fast’ (FSL 5.0.9) (Avants et al., 2008; RRID:SCR_002823; Tustison et al., 2010; Zhang et al., 2001). A T1w-reference map was computed after registration of the T1w images (after INU-correction) using ‘mri_robust_template’ (FreeSurfer 6.0.1) (Reuter et al., 2010). Brain surfaces were reconstructed using ‘recon-all’ (FreeSurfer 6.0.1) (Dale et al., 1999; RRID:SCR_001847; Reuter et al., 2010), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438; Klein et al., 2017). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with ‘antsRegistration’ (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) (RRID:SCR_008796, TemplateFlow ID: MN152NLin2009cAsym). Where available, T2w-images were included for surface reconstruction.

6.1.1.2 Functional data preprocessing

For each BOLD run, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. The BOLD reference was then co-registered to the T1w reference using ‘bbregister’ (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using ‘mcflirt’ (FSL 5.0.9) (Jenkinson et al., 2002). The BOLD time-series (including slice-timing correction) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by (Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks.

All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using ‘antsApplyTransforms’ (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Many internal operations of fMRIPrep use Nilearn 0.5.2 (RRID:SCR_001362; Abraham et al., 2014), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep’s documentation.

---

6 The description in this section was automatically created by fMRIPrep and adapted where needed.

7 https://fmriprep.readthedocs.io/en/latest/workflows.html
6.1.2 Deviation from preregistration

This analysis has been preregistered (Liem et al., 2019). While we have largely followed the plan, the analysis deviates in several minor points:

- Subjects were only considered for the study if they had at least three clinical sessions (not two as preregistered). Extracting the slopes of cognitive decline from two sessions resulted in very noisy slopes. As a result, the sample size is not 849 as preregistered, but 662. As our learning curve experiments demonstrate, the resulting sample size is sufficient in the present context.

- After further discussion, we added two features that have not been preregistered to the non-brain feature set: i) the number of sessions prior to the baseline session (to account for retest effects), and ii) the cognitive diagnosis at baseline (healthy, mild cognitive impairment, dementia). Both features did not show high importance in the permutation importance analysis.

- The preregistered structural features set included 331 features from global, subcortical, and cortical (volume and thickness) markers. After finding that the reduced structGS feature set (35 global and subcortical markers) performs equally well, we conducted the analyses with the more parsimonious structGS modality.

- After further discussion, the preregistered dimensionality reduction approach for the functional connectivity data seemed suboptimal. It averages positive and negative values, which might result in the cancellation of positive and negative connectivity within a network. The updated approach downsamples the connectivity matrix using a PCA.

- In the preregistration, two hyperparameters were planned to be used for tree pruning (max_depth = [5, 10, 20, 40, 50, None], where None leads to fully grown trees and min_samples_leaf = [1, 4, 10]). After further discussion, we decided to remove this double parametrization and only tune max_depth. Additionally, a currently ongoing independent project suggested that lower values of max_depth might be worth investigating in more detail. Hence, we added, max_depth = [3, 7, 15] to the hyperparameter tuning. Since the hyperparameters did not have a large influence on the result and those added values did not render the best performance we deem this change inconsequential.

- The results of the main analysis suggested further validation analyses (brain-age, extreme group classification). Those were not preregistered. However, they are very similar to the main analysis (only the models were adapted to the question: ridge regression for the brain-age analysis; random forest classifier for the extreme group classification with the criterion to measure the quality of an RF-split tuned with [“gini”, “entropy”]).
### 6.1.3 List of features

| Modality | # features | Name                          | Test/Variables                        | Information                                                                 | # features |
|----------|------------|-------------------------------|---------------------------------------|-----------------------------------------------------------------------------|------------|
| non-brain|            | demographic information      | Age, sex, education                   |                                                                             | 3          |
|          | 66         | clinical assessments         | MMSE                                  | Sum score                                                                   | 1          |
|          |            |                               | CDR                                   | 6 items, global and SOB scores                                              | 8          |
|          |            |                               | FAQ                                   | 10 items and sum score                                                      | 11         |
|          |            |                               | NPI-Q                                 | Symptom presence and severity sum scores                                    | 2          |
|          |            |                               | GDS                                   | Sum score                                                                   | 1          |
|          |            | neuropsychology               | WMS-R                                 | Logical memory (LOGIMEM, MEMUNIT, MEMTIME)                                  | 7          |
|          |            |                               |                                       | Digit span (DIGIF, DIGIFLEN, DIGI, DIGILEN)                                  |            |
|          |            |                               | Word fluency                          | ANIMALS, VEG                                                                | 2          |
|          |            |                               | TMT                                   | Part A (TRAILA, TRAILARR, TRAILAL), Part B (TRAILB, TRAILBRR, TRAILBLI),    | 7          |
|          |            |                               |                                       | TRAILBnorm = TRAILB/TRAILA                                                  |            |
|          |            |                               | WAIS-R                                | Digit Symbol                                                                | 1          |
|          |            |                               | BNT                                   | BOSTON                                                                      | 1          |
|          |            | APOE                          | ε 2, ε 3, and ε 4 allele count         |                                                                             | 3          |
|          |            | cognitive diagnosis           | healthy control, MCI, or dementia     |                                                                             | 1          |
|          |            | health                        | cardio/cerebro-vascular health, diabetes, hypercholesterolemia, smoking, and family history of dementia | 17         |
|          |            | N sessions before baseline    |                                       |                                                                             | 1          |
| structGS | 35         | subcortical volume            | accumbens, amygdala, caudate, hippocampus, pallidum, putamen, thalamus       | Volume of left and right                                                    | 14         |
|          |            | global measurements           | L + R mean cortical thickness; L + R lateral, 3rd, 4th ventricles; L + R total cortical volume; L + R cerebral white matter volume; L + R cerebellar white matter and cortical volume; total subcortical gray matter volume, total gray matter volume, corpus callosum volume (5 parcels) | 21         |
| func     | 100        | functional connectivity      | 300 cortical, cerebellar, and subcortical ROIs                                | reduced to 100 PCA components                                               | 100        |

Table S1. Input features
6.2 Supplementary Results

6.2.1 Predictive targets

Figure S1. Distribution of predictive targets (cognitive slopes of CDR and MMSE).

6.2.2 Predictive performance (full results)

Figure S2. Adding structural data to non-brain data improves prediction of cognitive decline. Test performance ($R^2$, coefficient of determination) across splits. Targets: cognitive decline measured via CDR (Clinical Dementia Rating, top) and MMSE (Mini-Mental State Examination, bottom). Input modalities: non-brain, structGS (global and subcortical structural volumes), struct (structGS + cortical volume and thickness), func (functional connectivity). Left panel represents combinations of input modalities (e.g., first line 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). This figure is an extension of Figure 2 and also includes brain modalities on their own.
### Table S2. Comparison of test performance vs non-brain predictions (% of splits for which the test prediction $R^2$ is outperforming the non-brain prediction; $N_{splits} = 1000$). Mean and SD difference show the mean and standard deviation of the modality’s performance vs. the non-brain prediction. Best performing model (gray): the CDR prediction from non-brain + structGS is outperforming the non-brain prediction in 91% of splits.

| modality / target          | outperforming non-brain (% of splits) | mean difference | SD difference |
|----------------------------|---------------------------------------|-----------------|---------------|
|                            | CDR change | MMSE change | CDR change | MMSE change | CDR change | MMSE change |
| structGS                  | 19         | 5           | -0.09      | -0.13       | 0.1        | 0.09        |
| struct                     | 17         | 7           | -0.09      | -0.12       | 0.1        | 0.09        |
| func                       | 0          | 0           | -0.36      | -0.31       | 0.08       | 0.08        |
| non-brain + structGS      | 92         | 78          | 0.06       | 0.02        | 0.04       | 0.04        |
| non-brain + struct         | 75         | 59          | 0.04       | 0.01        | 0.06       | 0.05        |
| non-brain + func           | 25         | 31          | -0.02      | -0.01       | 0.03       | 0.03        |
| non-brain + structGS + func| 78         | 57          | 0.04       | 0.01        | 0.05       | 0.04        |

### Table S3. Comparison of test performance vs null-models (% of splits for which the test prediction $R^2$ is outperforming the null-model prediction). Best performing model (gray): the CDR prediction from non-brain + structGS is outperforming the non-brain prediction in 100% of splits.

| modality / target          | outperforming null-models (% of splits) | median difference | SD difference |
|----------------------------|-----------------------------------------|-------------------|---------------|
|                            | CDR change | MMSE change | CDR change | MMSE change | CDR change | MMSE change |
| non-brain                  | 100        | 100         | 0.42       | 0.36        | 0.09       | 0.09        |
| structGS                   | 100        | 99          | 0.31       | 0.21        | 0.08       | 0.07        |
| struct                     | 100        | 100         | 0.34       | 0.24        | 0.08       | 0.07        |
| func                       | 91         | 73          | 0.01       | 0.01        | 0.01       | 0.01        |
| non-brain + structGS       | 100        | 100         | 0.48       | 0.39        | 0.08       | 0.09        |
| non-brain + struct         | 100        | 100         | 0.47       | 0.38        | 0.09       | 0.09        |
| non-brain + func           | 100        | 100         | 0.42       | 0.35        | 0.08       | 0.07        |
| non-brain + structGS + func| 100        | 100         | 0.47       | 0.37        | 0.08       | 0.08        |
Figure S3. Scatter plot of true vs predicted trajectories of cognitive change for the non-brain + structGS model. Mean prediction across splits. CDR: positive values = cognitive decline, MMSE: negative values = cognitive decline.
6.2.3 Hyperparameter tuning of the random forest regression model

Figure S4. Tuning curves of the random forest regression hyperparameters max tree depth (0 represents fully grown trees) and criterion (MSE: mean squared error, MAE: mean absolute error). Note that the y-axis is trimmed and the results vary in a rather narrow range.
6.2.4 Learning curves demonstrate sufficient sample size

Figure S5. Learning curve plotting the models’ performance across an increasing training sample size.
6.2.5 Age prediction

Figure S6. Predictive performance in 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. This figure is an extension of Figure 4 and also includes brain modalities on their own.

6.2.6 Extreme group prediction

Figure S7. Predictive performance in predicting extreme groups of cognitive decline (stable vs decline; F1-score: harmonic mean of the precision and recall). 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.
C. Barrett, R. Holsheder, W. Jessel, M. Meneghini, L. Pulgaron ER, 
Delafram AM, Suzuki S-I, Kunias Y, Denollet J. 2013. 
Trail-Making Test In: Gellman MD, Turner JR, editors. 
Encyclopedia of Behavioral Medicine. New York, NY: 
Springer New York. pp. 1986–1987.
doi:10.1007/978-1-4419-1005-9_138

Hojati SH, Ebrahimzadeh A, Khaeze A, Babajani-Feremi A, 
Alzheimer's Disease Neuroimaging Initiative. 2018. Predicting 
conversion from MCI to AD by integrating rs-fMRI and structural 
MRI. Comput Med Biol Med 102 30–39.
doi:10.1016/j.cobmed.2016.09.004

Jenkinson M, Bannister P, Brady M, Smith S. 2002. Improved 
Optimization for the Robust and Accurate Linear Registration 
and Motion Correction of Brain Images. Neuroimage 17:825–841.
doi:10.1016/S1053-8119(02)01132-7

Jette AM, Davies AR, Cleary PD, Calkins DR, Rubenstein LV, Fink A, 
Kosecoff J, Young RT, Brook RH, Delbanco TL. 1986. The 
Functional Status Questionnaire: reliability and validity when 
used in primary care. J Gen Intern Med 1:143–149.
doi:10.1007/978-1-4419-1005-9_138

Klein A, Ghosh SS, Bao FS, Giard J, Häme Y, Stavsky E, Lee N, Rossa 
Jette AM, Davies AR, Cleary PD, Calkins DR, Rubenstein LV, Fink A, 
Kosecoff J, Young RT, Brook RH, Delbanco TL. 1986. The 
Functional Status Questionnaire: reliability and validity when 
used in primary care. J Gen Intern Med 1:143–149.
doi:10.1007/978-1-4419-1005-9_138

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, 
Rosso D, Lazaro-Paz D, Delalleau O, Paragios N. 2011. Scikit-learn: 
Machine Learning in Python. J Mach Learn Res 12:2825–2830.
doi:10.1007/978-1-4419-1005-9_138

Piper J, Bal MM, Wibber J, Laubert A, Adelberg J, Haines J, 
Veit A, Yehle J, Murphy B, Weisberg RG. 2010. The 
Neurosynth database: a comprehensive resource for neuroimaging 
research. Neuroimage 49:1019–1025. doi:10.1016/j.neuroimage.2009.05.010

Rahim M, Thirion B, Varoquaux G. 2017. Joint 
prediction of multiple scores captures better individual traits from 
brain images. Neuroimage 158:145–154.
doi:10.1016/j.neuroimage.2017.06.072

Rahim M, Thirion B, Comtat C, Varoquaux G, Alzheimer’s Disease 
Neuroimaging Initiative. 2016. Translational Learning of 
Functional Networks for Alzheimer’s Disease Prediction. IEEE J 
Sel Top Signal Process 10:120–1213.
doi:10.1109/JSTSP.2016.2600400

Rahim M, Thirion B, Varoquaux G. 2019. Population shrinkage of 
covariance (PCoE) for better individual brain 
functional-connectivity estimation. Med Image Anal 54:138–148.
doi:10.1016/j.media.2019.03.001

Rathore S, Habes M, Ifikhar MA, Shacklett A, Davatzikos C. 2017. A 
review on neuroimaging-based classification studies and 
associated feature extraction methods for Alzheimer’s disease 
and its prodromal stages. Neuroimage 158:530–548.
doi:10.1016/j.neuroimage.2017.03.057

Reuter M, Rosas HD, Fischl B. 2010. Highly accurate inverse 
consistent registration: A robust approach. Neuroimage 53:1181–1196. 
doi:10.1016/j.neuroimage.2010.07.020

Scheinost D, Noble S, Horien C, Greene AS, Lake EM, Salehi M, Gao S, 
Shen X, O’Connor D, Barron DS, Yip SW, Rosenberg MD, 
Knable RT. 2019. Ten simple rules for predictive modeling of 
individual differences in neuroimaging. Neuroimage 193:35–45.
doi:10.1016/j.neuroimage.2019.02.057

Seabold S, Perktold J. 2010. Statsmodels: Econometric and statistical 
modeling with pythonProceedings of the 9th Python in Science Conference. Scipy. p. 61.
doi:10.1016/j.neuroimage.2017.03.057

Seltzman BA, Gratton C, Marek S, Raut RV, Dosenbach NU, 
Schlagal BL, Petersen SE, Di Girolamo A. 2015. A set of 
functionally-defined brain regions with improved representation of the 
subcortex and cerebellum. Neuroscience.
doi:10.1016/j.neuroimage.2017.03.057

Skock M, Collins J, Callahan-Flintoff C, Bowman H, Wible B. 2016. I 
GRIED A BUNCH OF THINGS! THE DANGERS OF UNEXPECTED 
OVERFITTING IN CLASSIFICATION. Neuroscience.
doi:10.1016/j.gis.2015.09.005

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, 
Gee JC. 2010. N4ITK: Improved N3 Bias Correction. IEEE Trans 
Med Imaging 29:1310–1320. doi:10.1109/TMI.2010.2046908

Varoquaux G. 2018. Cross-validation failure: Small sample sizes 
lead to large errors. Bar Crone. Neuroimage 180:69–77.
doi:10.1016/j.neuroimage.2017.06.061

Vogel JW, Vachon-Presseau E, Pitchi Binette A, Tam A, Orban P, La 
Joie R, Savard M, Picard C, Poirier J, Belle P, Breitner JCS, 
Villeneuve S, Alzheimer’s Disease Neuroimaging Initiative*. 
and the PREVENT-AD Research Group. 2018. Brain properties 
predict proximity to symptom onset in sporadic Alzheimer’s disease. 
Brain 141:1871–1883. doi:10.1093/brain/awy993

Wachinger C, Salat DH, Weimer M, Reuter M, Alzheimer’s Disease 
Neuroimaging Initiative. 2016. Whole-brain analysis reveals 
increased neuroanatomical asymmetries in dementia for 
hippocampus and amygdala. Brain 193:3252–3266.
doi:10.1093/brain/awv243

Wintraub S, Salmon D, Mercaldo N, Ferris S, Graff-Radford NR, Chui 
H, Crum practices in primary care. J Gen Intern Med 12:233–239. doi:10.1176/jnp.12.2.233

Woo C-W, Chang LJ, Lindquist MA, Wager TD. 2017. Building better 
models: biomarkers in translational neuroimaging. Nat 
Neurosci 20:365–377. doi:10.1038/s41593-017-0046-y

Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, Leirer VO. 
1982. Development and validation of a geriatric depression 
screening scale: a preliminary report. J Psychiatr Res 17:37–49.
doi:10.1016/0022-3956(82)90001-9

Zhang Y, Brady M, Smith S. 2001. Segmentation of the whole brain 
and subfields using multiple automatically generated templates. Neuroimage 101:495–512.