# Predicting brain-age from multimodal imaging data captures cognitive impairment

Franziskus Liem[a,*], Gaël Varoquaux[i,j], Jana Kynast[b], Frauke Beyer[b,c], Shahrzad Kharabian Masouleh[b], Julia M. Huntenburg[a,e], Leonie Lampe[b,f], Mehdi Rahim[i,j], Alexandre Abraham[i,j], R. Cameron Craddock[k,l], Steffi Riedel-Heller[f,g], Tobias Luck[f,g], Markus Loeffler[f,h], Matthias L. Schroeter[b,f,d], Anja Veronica Witte[b,c,f], Arno Villringer[b,d,c,f], Daniel S. Margulies[a]

[a] *Max Planck Research Group for Neuroanatomy & Connectivity, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*
[b] *Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*
[c] *Subproject A1, Faculty of Medicine, Collaborative Research Centre 1052 "Obesity Mechanisms", University of Leipzig, Leipzig, Germany*
[d] *Clinic for Cognitive Neurology, University of Leipzig, Leipzig, Germany*
[e] *Neurocomputation and Neuroimaging Unit, Department of Education and Psychology, Free University of Berlin, Berlin, Germany*
[f] *Leipzig Research Center for Civilization Diseases (LIFE), University of Leipzig, Leipzig, Germany*
[g] *Medical Faculty, Institute of Social Medicine, Occupational Health and Public Health (ISAP), University of Leipzig, Leipzig, Germany*
[h] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany*
[i] *Parietal project team - INRIA, Saclay, France*
[j] *CEA, DSV, I2BM, Neurospin, Gif-Sur-Yvette, France*
[k] *Computational Neuroimaging Lab, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA*
[l] *Center for the Developing Brain, Child Mind Institute, New York, NY, USA*

## Abstract

The disparity between the chronological age of an individual and their brain-age measured based on biological information has the potential to offer clinically-relevant biomarkers of neurological syndromes that emerge late in the lifespan. While prior brain-age prediction studies have relied exclusively on either structural or functional brain data, here we investigate how multimodal brain-imaging data improves age prediction. Using cortical anatomy and whole-brain functional connectivity on a large adult lifespan sample (N = 2354, age 19-82), we found that multimodal data improves brain-based age prediction, resulting in a mean absolute prediction error of 4.29 years. Furthermore, we found that the discrepancy between predicted age and chronological age captures cognitive impairment. Importantly, the brain-age measure was robust to confounding effects: head motion did not drive brain-based age prediction and our models generalized reasonably to an independent dataset acquired at a different site (N = 475). Generalization performance was increased by training models on a larger and more heterogeneous dataset. The robustness of multimodal brain-age prediction to confounds, generalizability across sites, and sensitivity to clinically-relevant impairments, suggests promising future application to the early prediction of neurocognitive disorders.

*Keywords:* Machine learning, Head motion, Cognition, Biomarker

## Highlights

- Brain-based age prediction is improved with multimodal neuroimaging data.
- Participants with cognitive impairment show increased brain aging.
- Age prediction models are robust to motion and generalize to independent datasets from other sites.

*Corresponding author
*Email addresses:* liem@cbs.mpg.de (Franziskus Liem), gael.varoquaux@inria.fr (Gaël Varoquaux), kynast@cbs.mpg.de (Jana Kynast), fbeyer@cbs.mpg.de (Frauke Beyer), kharabian@cbs.mpg.de (Shahrzad Kharabian Masouleh), huntenburg@cbs.mpg.de (Julia M. Huntenburg), lampe@cbs.mpg.de (Leonie Lampe), rahim.mehdi@gmail.com (Mehdi Rahim), abraham.alexandre@gmail.com (Alexandre Abraham), cameron.craddock@childmind.org (R. Cameron Craddock), steffi.riedel-heller@medizin.uni-leipzig.de (Steffi Riedel-Heller), tobias.luck@medizin.uni-leipzig.de (Tobias Luck), markus.loeffler@imise.uni-leipzig.de (Markus Loeffler), schroet@cbs.mpg.de (Matthias L. Schroeter), witte@cbs.mpg.de (Anja Veronica Witte), villringer@cbs.mpg.de (Arno Villringer), margulies@cbs.mpg.de (Daniel S. Margulies)

## 1. Introduction

The brain continues to change throughout adult life. Structural aspects, such as cortical thinning, demonstrate robust patterns of alteration during adulthood (Hogstrom et al., 2013; Storsve et al., 2014). Likewise, age-related differences in brain function, demonstrated through studies of functional connectivity, have also been observed (Damoiseaux et al., 2008; Dennis & Thompson, 2014).

Establishing the trajectories of such changes over the lifespan provides a basis for characterizing clinically-relevant deviations (Ziegler et al., 2012; Raz & Rodrigue, 2006). Brain-based age prediction offers a promising approach for providing personalized biomarkers of future cognitive impairments by capturing deviations from typical development of brain structure and function.

Brain-based age prediction aims to estimate a person's age based on brain data acquired using magnetic resonance imaging (MRI, Franke et al., 2010; Franke & Gaser, 2012). In a first step, an age prediction model is trained based on brain imaging data from a large lifespan sample. In a second step, this model can be used to estimate a novel individual's age based solely on their brain-imaging data. By comparing a person's estimated age with their chronological age, conclusions about age-typical and atypical brain development can be drawn.

Brain-based age prediction exemplifies a larger trend in neuroscience (Bzdok, 2016; Gabrieli et al., 2015; Pereira et al., 2009; Varoquaux & Thirion, 2014) and psychology (Yarkoni & Westfall, 2016) to move from correlative to predictive studies, often using tools from machine learning. Individual brain-based prediction and classification may give rise to brain imaging-based biomarkers that could aid clinical diagnostics, for instance, by predicting an individual's risk of developing dementia based on their brain (Bron et al., 2015).

One successful age prediction framework is based on structural brain data analyzed with voxel-based morphometry (VBM, Franke et al., 2010; Franke & Gaser, 2012). Using this approach, accelerated brain aging was found in patients with Alzheimer's disease (Franke et al., 2010; Franke & Gaser, 2012), traumatic brain injuries (Cole et al., 2015), psychiatric disorders (Koutsouleris et al., 2013), and subjects with risks to physical health (Franke et al., 2014). This brain-age metric can also predict the future conversion from mild cognitive impairment to Alzheimer's disease (Gaser et al., 2013). This computational approach is not restricted to showing accelerated brain aging as a negative effect but has also been used to demonstrate the positive effects of education, physical exercise (Steffener et al., 2016), and meditation (Luders et al., 2016) on brain aging. Other work has shown that accelerated brain development is related to accelerated cognitive development in young subjects (Erus et al., 2014).

In addition to brain structure, functional connectivity based on resting-state fMRI data (Craddock et al., 2013) also has the potential to provide clinically-relevant biomarkers (Craddock et al., 2009; Castellanos et al., 2013), as the data is easily acquired in a clinical setting (Greicius, 2008; Damoiseaux et al., 2012). Similar to the structural age estimation approach, Dosenbach et al. (2010) demonstrated that this is also feasible with resting-state functional connectivity data from young subjects. As different MRI modalities capture not only shared but also unique information about brain aging (Groves et al., 2012), prediction accuracy may benefit by incorporating these additional sources of information. For instance, Brown et al. (2012) and Erus et al. (2014) have shown that combining information from gray and white matter anatomy increases prediction accuracy in young subjects. The present study investigates how combining data from two even more dissimilar sources, brain anatomy and functional connectivity, influences age prediction in a lifespan sample. This is important as function and structure convey converging as well as diverging information (Damoiseaux & Greicius, 2009).

While machine-learning methods enable predictions on a single-subject level, factors driving these predictions are often difficult to determine. Predictions that appear to be based on brain information may actually be driven by confounds. One major confound in functional and structural MRI is head motion (Satterthwaite et al., 2013; Power et al., 2012; Reuter et al., 2015; Alexander-Bloch et al., 2016). For instance, head motion can make cortex appear thinner (Reuter et al., 2015). An age-related increase in head motion might give rise to a supposedly 'brain-based' age predictor that relies heavily on head motion. Furthermore, while machine learning models are trained on one dataset and evaluated on another, in neuroimaging these datasets often come from the same study, i.e., same site and scanner. In such cases, models may overfit one site's subtle idiosyncrasies, rendering poor predictive power for data from another site. Therefore, in the current study we aimed to address these confounds by determining the effect of head motion on brain-based age prediction and predictive performance on data from a novel site.

The present study investigates (i) whether incorporating multiple imaging modalities increases prediction accuracy, (ii) whether cognitive impairments are related to brain aging, and (iii) how robust our predictive models are, specifically regarding head motion and generalizabilty to new datasets. Using data from brain anatomy and functional connectivity, we show that (i) incorporating multiple modalities increases predictive performance, (ii) cognitive impairments are related to advanced brain aging, and (iii) our models are robust as they are not driven by head motion and generalize reasonably to new datasets.

## 2. Materials: lifespan data & preprocessing

Two independent samples were investigated in this study: the LIFE (Loeffler et al., 2015) and the Enhanced Nathan Kline Institute – Rockland sample (NKI, Nooner et al., 2012). Since the majority of analyses is performed on the LIFE dataset, the NKI set is described in detail in Appendix A.

### *2.1. LIFE sample*

Participants took part in the LIFE-Adult-Study (`life.uni-leipzig.de`, Loeffler et al., 2015) of the Leipzig Re-

search Centre for Civilization Diseases (LIFE) and were randomly selected, community-dwelling volunteers. The study was approved by the institutional ethics board of the Medical Faculty of the University of Leipzig. Participants signed an informed consent form and were paid for their participation.

Of the 10000 volunteers participating in the LIFE-study, approximately 2600 also underwent MRI assessment and neuropsychological testing. In the present study, data from 2354 individuals were included. Exclusion criteria included neuroradiological findings, missing MRI data or excessive head motion in the functional scans (mean FD > 0.6 mm, Power et al., 2012). Subjects were between 19 and 82 years ($M = 58.68; SD = 15.17$; 1120 female, 1234 male).

### 2.2. Cognitive phenotyping

*Neurocognitive Disorder.* The Diagnostic Statistical Manual of Mental Disorders 5th edition (DSM-5, American Psychiatric Association, 2013) introduced Neurocognitive Disorder (NCD) as a new diagnostic category for acquired cognitive dysfunction. NCD diagnosis comprises the evaluation of subjective cognitive complaints, cognitive performance and independence in activities of daily living. To this end, a comprehensive and domain specific neuropsychological evaluation is required, including the cognitive domains attention, executive function, memory, language, visuoconstruction, and social cognition.

Two subtypes of NCD are distinguished: mild and major NCD. Both sub-types are characterized by subjective cognitive complaints. Mild NCD is presented with a cognitive performance decline that ranges between -1 and -2 SD below age and gender norms in at least one cognitive domain, and preserved independence in daily life. Contrary, persons with major NCD have severe cognitive deficits (<-2 SD below age and gender norms) in at least one cognitive domain that interfere with independence in everyday activities. Thus, the term major NCD represents the current concept of dementia.

The DSM-5 criteria for NCD diagnosis served as a template to characterize the study sample with respect to objective cognitive impairment (OCI). Only OCI was investigated in the present study. Independence in daily activities and subjective cognitive complaints were not considered as a criterion for cognitive phenotyping in this study as consensus questionnaires are still lacking.

*Neuropsychological assessment.* Cognitive performance was assessed with a set of standard neuropsychological tests, spanning several cognitive domains (Loeffler et al., 2015). All scores were carefully checked for missing values and plausibility. The Stroop test (Stroop, 1935; Treisman & Fearnley, 1969; Zysset et al., 2001; Schroeter et al., 2002), which quantifies executive functions, was administered. Social cognition was assessed with the Reading the Mind in the Eyes test (RMET, Cohen et al., 2001; Bölte, 2005), which quantifies the ability to infer mental states only from eye gaze. The CERAD-plus (Thalmann et al., 1997; Morris et al., 1989) is a dementia screening battery focusing on Alzheimer's disease. It includes tests of verbal and figural memory and learning (10-items word list, figure recall), language (Boston Naming Test, semantic and phonematic verbal fluency), and visuoconstrution (figure copy). This battery also includes the Mini Mental State Examination (MMSE, Folstein et al., 1975) as well as the Trail Making Test (TMT, Reitan, 1979), which measures visual attention and cognitive flexibility.

*Domain-specific scores.* For the domain-specific evaluation of cognitive performance, test scores were assigned to the cognitive domains proposed by the DSM-5 and aggregated (Beck et al., 2014).

Scores for the following domains were calculated:

- *attention* (TMT-A time to complete (TTC), TMT-A errors, Stroop neutral reaction time (RT), Stroop neutral % correct),
- *executive function* (TMT-B TTC / TMT-A TTC, Stroop incongruent TTC / Stroop neutral TTC),
- *memory* (CERAD word list (trial 1, 2, 3, total, delayed recall, recognition), CERAD figure delayed recall),
- *language* (Boston Naming Test, semantic fluency (animals), phonematic fluency (s-words)),
- *visuoconstruction* (CERAD figure copy), and
- *social cognition* (RMET).

*Objective cognitive impairment (OCI).* To translate neuropsychological measures into scores informative of cognitive performance independent of age, sub-test scores were z-standardized within the corresponding age- and sex group (18-39, 40-49, 50-59, 60-64, 65-69, 70-74, 75+ years). Where necessary, standardized scores were inverted so that a higher score reflects higher performance. If more than one sub-test measure per domain was available, sub-tests were averaged within domains. These domain-specific scores can be interpreted as the average performance of a person in a certain domain and a deviation from their peers' performance in that domain. Based on the domain-specific scores and in analogy to aforementioned NCD classification scheme, subjects were classified as *OCI-norm*, -*mild* (at least one domain score between -1 and -2 SD), or -*major* (at least one domain score <-2 SD), if they had at least three valid domain scores.

### 2.3. MR data

Brain imaging was performed using a 3T Siemens Trio scanner equipped with a 32 channel head coil.

Resting-state functional images were acquired using an T2*-weighted echo-planar imaging sequence with an in-plane voxel size of 3×3 mm, slice thickness of 4 mm, slice gap of 0.8 mm, 30 slices, echo time (TE) of 30 ms, repetition time (TR) of 2000 ms and a flip-angle of 90°. This sequence lasted 10 min (300 volumes), during which participants were instructed to keep their eyes open and not to fall asleep. A gradient-echo fieldmap with the same geometry was recorded for distortion correction (TR = 488 ms, TE 1 = 5.19 ms, TE 2 = 7.65 ms).

High resolution T1-weighted structural images were acquired using an MP-RAGE sequence with 1 mm isotropic voxels, 176 slices, TR = 2300 ms, TE = 2.98 ms, and inversion time (TI) = 900 ms.

### *2.4. MR data preprocessing*

MRI data processing was implemented in a python pipeline via Nipype (v0.10.0, Gorgolewski et al., 2011), which included routines from FSL (v5.0.9, Jenkinson et al., 2012), FreeSurfer (v5.3, Fischl, 2012), ANTS (v2.1.0, Avants et al., 2011), CPAC (v0.3.9.1, `fcp-indi.github.io`), and Nilearn (v0.2.3, `nilearn.github.io`, Abraham et al., 2014). The pipeline is available at `github.com/fliem/LIFE_RS_preprocessing`.

*Functional MRI.* After removal of the first five volumes (to allow the magnetization to reach a steady state) and motion correction (FSL mcflirt), rigid body coregistration of the functional scan to the anatomical image (FreeSurfer bbregister), as well as EPI distortion corrections (FSL fugue) were calculated and jointly applied in a subsequent step to each volume of the functional scan. Denoising included removal of (i) 24 motion parameters (CPAC, Friston et al., 1996), (ii) motion and signal intensity spikes (Nipype rapidart), (iii) six components explaining the highest variance from a singular value decomposition of white matter and cerebrospinal fluid time series (CompCor, Behzadi et al., 2007, signals extracted from individual masks created with FSL fast, decomposition executed with CPAC), and (iv) linear and quadratic signal trends. Subsequently, functional data were morphed to MNI space via transformation fields estimated from the structural data (ANTS). Functional data were then band-pass filtered between 0.01 and 0.1 Hz (Nilearn).

*Structural MRI.* The FreeSurfer software package was used to create models of the cortical surface for cortical thickness and cortical surface area measurements. Subcortical volumes were obtained from the automated procedure for volumetric measures of brain structures implemented in FreeSurfer.

## 3. Methods: age prediction analysis

### *3.1. Age prediction*

Models were trained to predict age based on a variety of input data, i.e., functional connectomes of two different spatial resolutions and measures of cortical anatomy (cortical thickness, cortical surface area, subcortical volumes). A schematic overview of the age prediction analysis is shown in Figure 1.

#### *3.1.1. Input data*

The following five sources of neuroimaging data entered the age prediction models. Two sources represent brain connectivity in different spatial resolutions, three sources originate from brain anatomy:

1. *connectivity matrix 197*,
2. *connectivity matrix 444*,
3. *cortical thickness*,
4. *cortical surface area*, and
5. *subcortical volumes*

After extracting feature vectors for each subject and modality (see Figure 1.1), vectors were stacked to obtain the input data matrices for the age prediction analysis (see Figure 1.2).

*Brain function.* Functional connectomes were derived from preprocessed functional MRI data using the Nilearn package. Mean time-series were extracted from cortical and subcortical regions of the functionally defined BASC parcellation atlas (Bellec et al., 2010, obtained via the Nilearn data fetcher fetch_atlas_basc_multiscale_2015). Functional connectivity between all pairs of regions was quantified via Pearson correlation, resulting in a symmetric connectivity matrix. Since measures derived from connectomes vary with parcellation resolution (Fornito et al., 2010) and there is no 'right' number of parcels, we investigated two different levels of spatial granularity. Based on Thirion et al. (2014), who recommend parcellations consisting of around 200 to 500 regions, we reconstructed connectivity matrices from 197 and 444 regions.

Connectivity matrices underwent Fisher's r-to-z transformation and a feature vector was extracted from the lower triangle ($N_{features}(connectivity\ matrix\ 197) = 19306$, $N_{features}(connectivity\ matrix\ 444) = 98346$). The shape of the input matrix was $N_{subjects} \times N_{features}$ (with $N_{subjects}$ varying between analyzes; see section 3.2).

*Brain anatomy.* Native surface models for cortical thickness and surface area were transformed into the fsaverage4 standard space. The data for the two hemispheres was concatenated ($N_{features}(cortical\ thickness) = N_{features}(cortical\ surface\ area) = 5124$). Volumetric data for subcortical regions and measures of global volume were extracted from the aseg.stats file ($N_{features}(subcortical) = 66$).

#### *3.1.2. Predictive analysis*

Predictive models were implemented in a two-level approach (see Figure 1.4). On the first level, linear support vector regression models (SVR, Drucker et al., 1996) were used to predict age from neuroimaging data (single-source models). On the second level, predictions from the single-source models were stacked with random forest (RF, Breiman, 2001) regression models. Using RF models for stacking multiple neuroimaging modalities has previously been shown to produce better predictions with smaller variability in prediction errors as compared to other stacking methods (Rahim et al., 2016). All predictive analyses have been performed using the python-based Scikit-learn package (Pedregosa et al., 2011). The code is available at `github.com/fliem/LeiCA_LIFE`.

In detail, this procedure entailed (see Figure 1):

1. *Train-test-split.* Data was split into equally sized training and test set (see Figure 1.3). The training set was used to train (learn) the models, the test set was put aside to subsequently evaluate the models' performance on unseen data.

2. *Training of single-source models (see Figure 1.4).* Two parallel approaches have been used to train single-source models. First, using the neuroimaging data of the entire training set, single-source SVR models were fitted, resulting in one trained model per source. Second, in parallel, using the neuroimaging data of the training set, single-source SVR models were trained in a 5-fold cross-validation (CV) approach (see Figure 1.4.1). This was done to obtain unbiased CV-predictions (to be used in the following step).
3. *Training of multi-source models.* To aggregate information from multiple sources into one prediction, the previous step's CV-predictions were stacked (concatenated). A feature vector based on the single-source predictions was constructed. Based on this feature vector, the multi-source models were fitted, to obtain trained multi-source models (see Figure 1.4.2). This was done in three versions:
   (a) *stacked-function* which combined age predictions from *connectivity matrix 197* and *connectivity matrix 444*,
   (b) *stacked-anatomy* which combined age predictions from *cortical thickness, cortical surface area, and subcortical volumes*, and
   (c) *stacked-multimodal* which combined age predictions from all five single-source models.

   For instance, in the case of *stacked-multimodal*, each subject's feature vector consisted of the five age prediction values from the single-source models.
4. *Test the models by predicting age in new subjects.* The performance of the trained single- and multi-source models was tested with the neuroimaging data of the test set as input (see Figure 1.5). First, single-source predictions are calculated by using the trained model from step 2 (see Figure 1.5.1). Second, these predictions are stacked and fed into the trained model from step 3 (see Figure 1.5.2), to receive single- and multi-source test set predictions.
5. *Evaluate generalization performance.* Finally, the models' generalization performance can be assessed via the test set's absolute error (AE) of the age predictions (obtained in step 4) from chronological age. Additionally, the coefficient of determination (R2) is also reported.

*Statistical tests.* To compare models, non-parametric statistical tests were run on the absolute prediction errors, using the SciPy package (v0.17.0, `scipy.org`). Correction for multiple comparisons was done using the false discovery rate (FDR) procedure described by Benjamini & Hochberg (1995). Results were plotted with the Seaborn package (v0.7.0, Waskom et al., 2016).

*Tuning of hyperparameters.* For the single-source SVR models, tuning curves for the $C$ parameter were run on the training data. These curves showed a 'sweet spot' for the high dimensional neuroimaging input data (the connectivity matrices, cortical thickness and cortical surface area) around $C = 1e{-}3$ (for an example see Figure A.6). For the lower dimensional subcortical input data, the standard $C = 1$ performed well. All models were run with the default $\varepsilon = 0.1$. For the multi-source RF models, out-of-bag estimates were used to set the tree depth.

### 3.2. Analysis plan

A brief sketch of the different analyzes, tailored to the different research questions, follows here. Further details can be found in the results section.

#### 3.2.1. Prediction performance in multimodal data

Age predictions have been performed as described in section 3.1. The entire LIFE sample was split into equally sized training and test set.

#### 3.2.2. Brain aging in cognitively impaired subjects

For this analysis, the sample was reduced to subjects with a valid OCI score (see section 2.2). Models were trained on *OCI-norm* subjects only. The test set consisted of subjects from all OCI groups. A brain aging (BA) score was then calculated for each subject and each single- and multi-source model by subtracting $age_{chronological}$ from $age_{predicted}$. These BA scores were compared between OCI groups.

#### 3.2.3. Robustness against confounds

*Head motion.* The robustness of our approach against head motion was investigated with the models described in section 3.2.1. First, we did this using *motion regression*. On the group-level, head motion (mean FD derived from the functional scans) was regressed out of the input matrices in the single-source models for the entire training and entire test set separately. Mean FD was derived from the functional acquisition and used as a measure of head motion for the functional as well as the structural data. Second, as an alternative, *motion matching* was performed by creating a subsample of the test set that does not show an $age \times motion$ correlation. For direct comparison, an equally sized random sample was also drawn. These test samples were then used to evaluate the performance of the models trained on the full training sample (see 3.2.1).

*Generalization to new site.* In this step we investigated how the models generalize to data from a new site (different country, scanner, acquisition protocol, subjects). Age predictions were performed on data from the NKI set using models trained on LIFE data (one sample training; section 3.2.1). In a subsequent analysis, models were re-trained on a training set that combined the original LIFE training sample with a small number of subjects from the NKI set, to increase the generalizability of the predictive models (two sample training). Finally, to increase the training sample size, the one and two sample training was repeated, including the majority of the LIFE data into the training sample, not only the original training set (99 % of the entire LIFE sample; 1% was retained for a rough check of the models on LIFE data). This analysis will be referred to as *full LIFE sample training* approach.

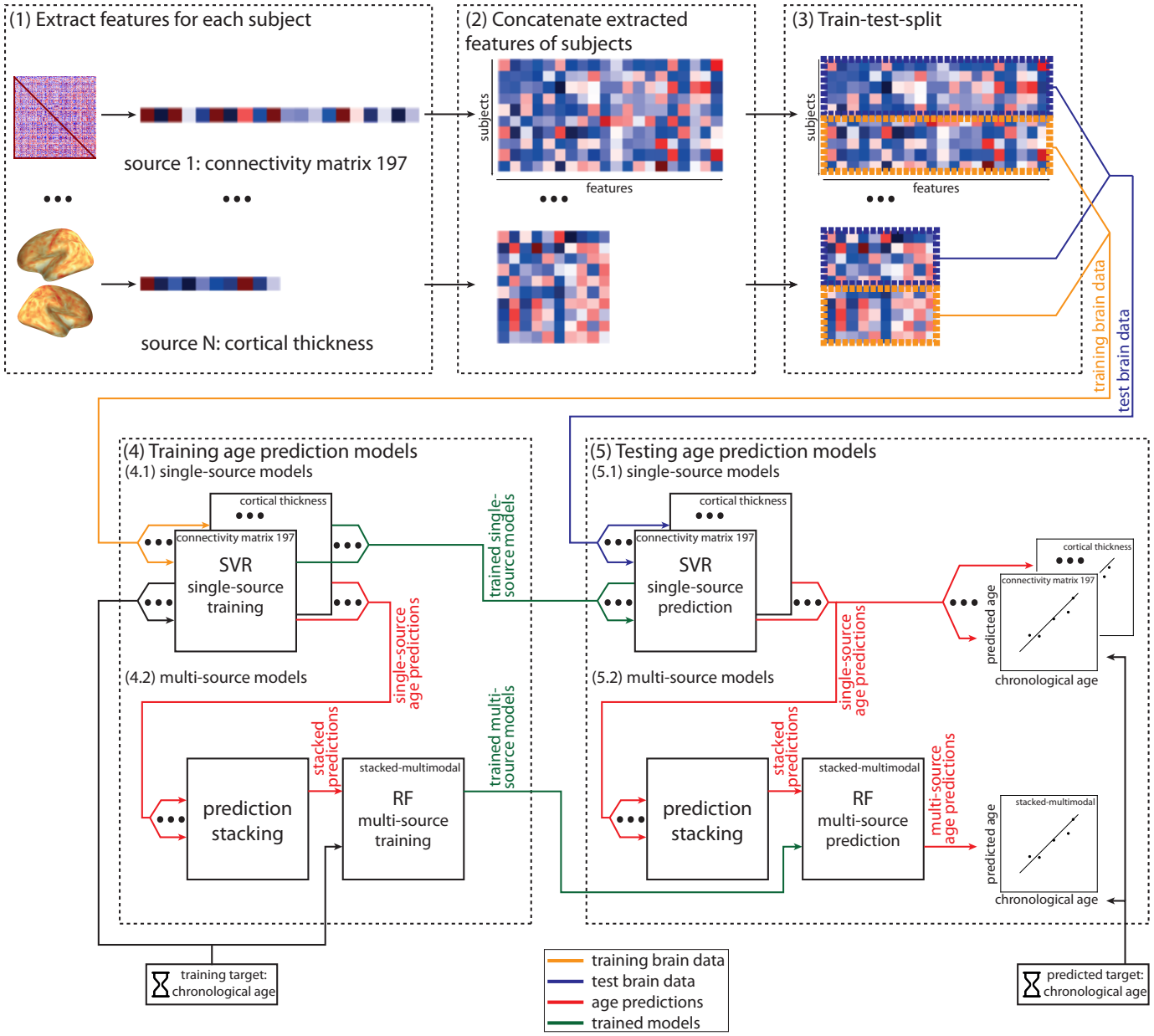


Figure 1: Overview of feature extraction and predictive analysis. (1) For each subject, feature vectors from the following data sources are extracted: *connectivity matrix 197, connectivity matrix 444, cortical thickness, cortical surface area*, and *subcortical volumes*. (2) Within each source, data for subjects are concatenated to obtain input data matrices. (3) Data is split into training and test set. (4) Training data (yellow line) is used to train age prediction models. (4.1) First, five single-source support vector regression models (SVR) are trained to predict chronological age based on training brain data. (4.2) Second, the single-source predictions (red line) are stacked and entered into the training of multi-source random forest models (RF). Three separate multi-source models were trained: *stacked-function* (combining *connectivity matrix 197* and *connectivity matrix 444*), *stacked-anatomy* (combining *cortical thickness, cortical surface area, and subcortical volumes*), and *stacked-multimodal* (combining all five single-source models). (5) The trained models (green line) are then evaluated. (5.1) Trained single-source models give single-source age predictions based on test data (blue line). (5.2) These predictions are stacked and entered into the trained multi-source models to obtain multi-source age predictions. Prediction performance is evaluated by comparing predicted age with chronological age.

## 4. Results

Our results demonstrate that i) incorporating multiple brain imaging modalities increases age prediction performance (Figure 2 and 3); ii) subjects with objective cognitive impairment show advanced brain aging compared to subjects without objective cognitive impairment (Figure 4); iii) our prediction models are robust against confounds (Figure 5), i.e., not driven by head motion and generalize to new datasets. For the comparison of modalities, data for models of all modalities will be presented. After showing that the multimodal approach outperforms the others, the remainder of the results, for the sake of brevity, will focus on this model. The full results can be found in Appendix B.

### *4.1. Multimodal data increases prediction performance*

Based on the entire LIFE sample, age prediction models were trained on the training set ($N = 1177$) and evaluated on the test set ($N = 1177$). Figure 2 shows prediction performance on the test sample (for full statistics see Table B.1). All models show good prediction performance (mean absolute error between 4.29 and 7.29 years, R2 between 0.62 and 0.87). The stacked models show better performance than single source models, with the *stacked-multimodal* model outperforming all other models. Additionally, this model also shows the least prediction variability. By going from the second best model, *stacked-anatomy*, to the best, *stacked-multimodal*, approximately half a year in prediction accuracy is gained. Table B.2 shows feature importances for the multi-source models.

Figure 3 shows the individual predictions for the *stacked-multimodal* model, the model with the best predictive performance.

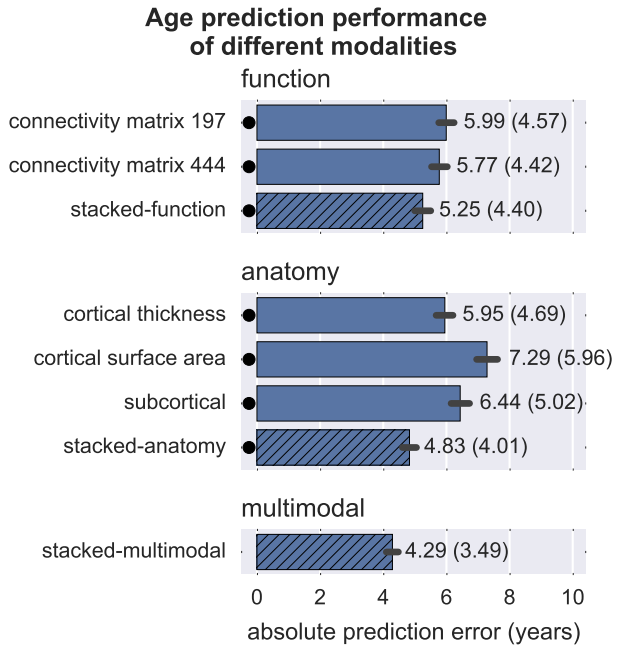


Figure 2: Prediction performance (absolute error on test set, lower values are better). Note that the *stacked-multimodal* model shows the least prediction error. Black dots represent significant ($p(FDR) < 0.05$) deviation from the best model, indicating that *stacked-multimodal* significantly outperforms all other models. Error bars represent 95% CI bootstrapped with 1000 iterations. Numbers next to error bars represent mean (standard deviation). Stacked models are shown with hatched bars. For full statistics see Table B.1

### *4.2. Advanced brain aging in cognitively impaired subjects*

Based on a large battery of cognitive tests, subjects with mild or major OCI were identified. For this analysis, the models were trained on half of the *OCI-norm* subjects ($N_{training} = 724$). Subsequently, age predictions were performed on a test sample containing *OCI-norm*, *mild* and *major* subjects (test: $N_{norm} = 729, N_{mild} = 632, N_{major} = 251$) and compared between groups (sample characteristics can be found in Table B.4). Figure 4 shows the advanced brain aging in OCI predicted by the *stacked-multimodal* model. Figure B.7 shows that all models, except *stacked-function*, show a significant progression in brain aging related to the severity of OCI (see Table B.6 for full statistics). This finding demonstrates that the age prediction models capture aspects of cognitive impairment.

### *4.3. Robustness against confounds*

*Head motion.* Our sample showed a substantial *age* × *motion* correlation ($r_{age \times motion} = 0.43$; $p = 1.87e{-}15$; head motion defined as mean FD derived from the functional scans). To test the influence of head motion on the age prediction models, the following two analyses have been performed.

*Motion regression.* To test the models' robustness against head motion, we regressed out head motion (mean FD) from the input data (for training and test set separately). Regressing out motion reduces prediction accuracy significantly (e.g., the *stacked-multimodal* model's error increases from 4.29 to 6.95 years; see Figure B.8 and Table B.7). This might either be the result of head motion driving the prediction models, or, due to the large variance shared by age, head motion and the brain measures, of too aggressively removing age-related variance while removing motion-related variance. To test these two alternative explanations, the following motion matching analysis was performed.

*Motion matching.* In this analysis, a motion adjusted subsample of the test set is created by restricting the sample to subjects with a mean FD between 0.19 and 0.28 mm and an age above 25 years, which results in a motion matched subsample ($N = 387$; $r_{age \times motion} = 0.06$; $p = 0.26$). Excluding subjects with a mean FD lower than 0.19 was necessary to create a balanced sample, because of the dominance of young subjects with low motion. An equally sized non-motion-adjusted subsample was randomly drawn for comparison ($r_{age \times motion} = 0.42$; $p = 2.45e{-}18$). These subsamples ($N = 387$) were used to evaluate the influence of motion in the models trained on the original training set ($N = 1177$; see 4.1). The *stacked-multimodal* model (Figure 5.1), as well as all other models (Figure B.9) perform equally well with and without motion matching (all $p > 0.49$; for full statistics see Table B.8), indicating that head

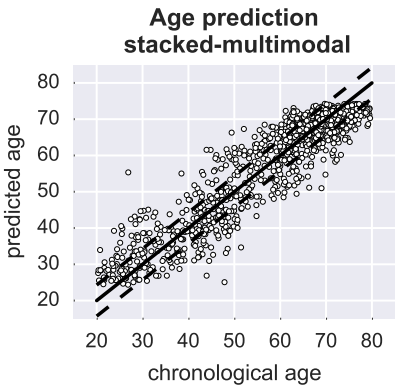


Figure 3: Chronological and predicted age from the *stacked-multimodal* model. Circles represent subjects, the solid line the perfect prediction, dashed lines the mean absolute prediction error (4.29 years).

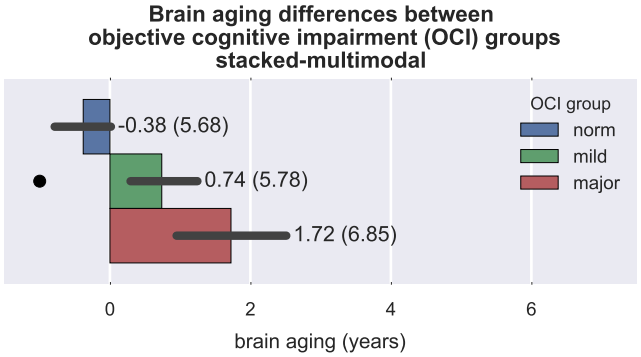


Figure 4: Differences in brain aging (brain aging = predicted age - chronological age) between OCI groups for *stacked-multimodal*. Positive brain aging values indicate that a brain appears older than expected from chronological age. Note that brain aging significantly increases with severity of OCI, i.e., more advanced brain aging in OCI. Numbers next to error bars represent mean and standard deviation. For full data see Figure B.7 and Table B.6.

motion is not driving the age prediction models and that motion regression is removing too much meaningful age-related variance.

*Generalization to new site.* To demonstrate how the models generalize to data from a new site (different country, scanner, acquisition protocol, subjects), we predicted age on NKI data with models that have been trained on LIFE data (one sample training). While the models perform much better than chance, unsurprisingly, better predictive performance is achieved on LIFE than on NKI data (Figure 5.2; for full data see Figure B.10 and Table B.9). Assuming that models show higher generalizabilty if trained on more heterogeneous data, the following post-hoc analysis tested whether adding a small number of subjects from NKI to the LIFE training sample increases generalization (two sample training; training sample: $N_{LIFE} = 1177; N_{NKI} = 46$, representing around 10% of the NKI sample).

While prediction performance increases by adding subjects


Figure 5: Robustness of age prediction against confounds for the *stacked-multimodal* model. (1) Motion matching analysis show that age prediction works equally good in motion adjusted (without *age* × *motion* correlation) and non-adjusted (with *age* × *motion* correlation) groups, indicating that the predictive model is not driven by head motion. Full data are shown in Figure B.9 and Table B.8. Note that the slightly lower prediction error (around 4.06) as compared to the original analysis (around 4.29; see Figure 2) is a result of the restricted age range of the test samples in the motion matching analysis. Hence, those values should only be compared within the motion matching analysis and not with the original analysis. (2) Generalization to new site. Standard training procedure (one sample training) showed significantly ($p(FDR) < 0.05$ as indicated by the black dot) better prediction performance in LIFE data (within site) than in NKI data (between site, for full data see Figure B.10 and Table B.9). (3) After training the model on a mixed-site sample (two sample training, $N_{LIFE} = 1177; N_{NKI} = 46$), predictions on the NKI data improve (Table B.10), but the predictions on the main training site LIFE (within site) still are significantly better than on the minor training site NKI (between site, for full data see Figure B.11 and Table B.11). (4) Finally, generalization is investigated by training on the full LIFE sample ($N_{training,LIFE} = 2377$). Test prediction performance on between-site NKI data for one sample training (LIFE sample only) and two samples (LIFE + NKI samples; $N_{training,NKI} = 46$) slightly increases as compared to the original training approach (green bars from (2) and (3); for full data see Figure B.12 and Table B.12).

from NKI to the training sample (see Table B.10), Figure 5.3 shows that prediction still works better on LIFE data (for full data see Figure B.11 and Table B.11). We also demonstrate that these results are robust across different random splits of the data (see Table B.13).

As a further attempt to increase generalizability, we pursued the *full LIFE sample training* approach. Here, we repeated the one and two sample training using the majority of all LIFE data for training (training samples: one sample training: $N_{LIFE}$ = 2377); two sample training. $N_{LIFE}$ = 2377; $N_{NKI}$ = 46). These trained models were then evaluated with the (remaining) NKI data. This further increases generalizability and reduced the prediction error. For the *stacked-multimodal* (two sample) analysis it decreases to 6.56 years (Figure 5.4), which is a slight reduction compared to the original two sample result of 6.93 years (for full data see Table B.12 and Figure B.12, for a scatter plot of test predictions Figure B.13).

## 5. Discussion

The aim of the current study was to establish a novel multimodal brain-based age prediction framework that makes use of information from anatomy and functional connectivity. We found that (i) including multimodal information increases prediction accuracy, (ii) objective cognitive impairment is associated with increased brain aging, and (iii) our framework is robust against confounds, most importantly, against head motion, and generalizes to new datasets, especially if the training set is composed of a large and heterogeneous dataset.

Age prediction was best achieved using the multimodal approach (*stacked-multimodal*), which showed a mean absolute age prediction error of 4.29 years. This is approximately a half-year more accurate than when only taking anatomical information into account (*stacked-anatomy*). Furthermore, the multimodal approach shows less variability in prediction performance. We assume that the gain in prediction accuracy is a result of the different brain-imaging modalities' shared variance, via reducing the measurement error of brain data, as well as unique variance, via the addition of new information. Aggregating multiple sources of neuroimaging data via Random Forest models has been shown to work well (Rahim et al., 2016). In particular, aggregating data via RF models results in better age prediction performance as compared to merely averaging single-source predictions (e.g., for *stacked-multimodal*: age prediction error of 4.29 vs 5.08 years).

Our anatomical approach is conceptually similar to the framework of Franke et al. (2010). The main difference is the choice of anatomical data analysis tool: voxel-based morphometry in their work, surface-based morphometry in ours. Their best model showed a mean absolute prediction error of 4.61, which is in agreement with the performance of *stacked-anatomy* at 4.83 years. The surface-based morphometry approach has the advantage of disentangling structural information of cortical thickness and surface area (Meyer et al., 2014). Age prediction based on cortical thickness worked better than based on cortical surface area, which is well in line with stronger age-related effects in cortical thickness than in surface area (Hogstrom et al., 2013). Future studies might also investigate whether considering additional information about white matter anatomy further reduces prediction accuracy. How much further the prediction accuracy can be reduced, i.e., the lower bound, is unclear. Due to individual differences in the brains of individuals of the same age, some prediction error will always persist.

We investigated brain aging in individuals with objective cognitive impairment. By subtracting chronological age from predicted age, we calculated a brain aging score (also called *brainAGE (brain age gap estimate)* by Franke et al. (2010), or *PAD (predicted age difference)* by Cole et al. (2015)). The multimodal approach, as well as most other approaches we investigated, predicted significantly increased brain aging in participants with objective cognitive impairment. The progression of brain aging always followed the progression of OCI and increased from normal to mild to major OCI individuals. The strongest differences in brain aging between the OCI groups was observed in the model using subcortical data. This suggests that while the multimodal approach performed best in age prediction, differences in cognitive performance might be better characterized using specific modalities. As different pathologies might be detectable early in different MRI modalities, future studies should consider the effectiveness of predictive models of different uni- and multimodal approaches in the context of a given pathology.

The brain-age metric provides an interpretable aggregate measure of brain aging processes in brain structure and function. However, if the primary research interest is predicting cognitive performance, why investigate this via metrics of brain-age? Ideally, the predictive model should be created using a study-specific cognitive target (for instance, see Ullman et al., 2014). Directly predicting future cognitive performance certainly holds tremendous potential to identify specific cognitive modalities at risk of future decline. These models offer a valuable foundation for innovating tailored interventions through, for example, cognitive training.

However, to obtain stable models large datasets with brain and behavioral data is required. Assessment of brain-age offers an alternative and complementary measure that is already available through several publicly available large-scale brain-imaging datasets. Such data can be used to train models, which are then complemented by a smaller, but richer dataset that includes information about cognitive performance, in order to test specific hypotheses.

Confounding effects of head motion in brain-imaging studies have received increased interest in the recent years (e.g., see Power et al., 2012; Reuter et al., 2015). The present study demonstrated that head motion does not drive brain-based age prediction and that regressing out motion might also affect meaningful age-related variance. While the estimation of head motion in functional MRI scans is well established, this is much more challenging for structural MRI scans due to their longer acquisition times. While there exist special acquisition protocols tailored to measure head motion (for instance see Reuter et al., 2015), these are not yet standard, and do not apply to already existing data. However, since head motion has within-

subject stability (Van Dijk et al., 2012), we took head motion estimates based on functional scans as a proxy for head motion in structural scans, as also done by Alexander-Bloch et al. (2016). Nevertheless, motion between different scan blocks certainly can differ, which might render the motion metrics derived from functional scans a poor proxy for structural scans. For instance, the time point of collecting the structural data (at the beginning or end of a scanning session) might result in different motion characteristics due to fatigue of the study participants or adaptation to the in-scanner situation. These effects have not yet been studied systematically and deserve attention in future studies.

Age prediction models perform significantly better when trained and tested on data from the same site, as compared to data from different sites. Training models on a larger and more heterogeneous dataset modestly improves the prediction accuracy. However, since even in this case within-site prediction outperforms between-site prediction this topic deserves more attention in future work. Several factors may have contributed to the better generalizability observed in the NKI dataset using anatomical rather than functional information. First, the anatomical sequences used in both studies are quite similar, while the functional sequences differ with regards to temporal and spatial parameters. Second, anatomical information analyzed with surface-based morphometry shows higher reliability (Liem et al., 2015) than functional MRI (Shehzad et al., 2009). To avoid fitting models to the idiosyncrasies of a given study, future studies should broaden the variability of training data by including data from an array of sites, as recently demonstrated (Abraham et al., 2016; Cole et al., 2015).

A standardization of MRI acquisition protocols may also contribute to a better generalization of predictive models. Quantitative structural MRI (Lutti et al., 2014) or calibration of functional sequences (Chiarelli et al., 2007) may provide more reliable and valid brain measurements, resulting in better predictors. However, these techniques are not currently standard practice and require further development for widespread application (e.g., see Dubois & Adolphs, 2016).

By moving from correlative studies to predictive studies using tools from machine learning (Gabrieli et al., 2015; Yarkoni & Westfall, 2016), cognitive neuroscience as a basic science might be complemented with an applied component that can give relevant insights into both clinical pathologies as well as the healthy spectrum of aging. This may range from brain-based biomarkers for neurological or psychiatric diseases, to identifying potential future cognitive impairments on an individual-level and designing targeted cognitive training.

## 6. Conclusions

In the present study, we demonstrated that including information from multiple MR modalities, i.e., anatomy and functional connectivity, increased accuracy of brain-based age prediction. Brain-age measured with this multimodal framework was accelerated in subjects with cognitive impairment. Importantly, head motion does not drive brain-based age prediction and predictive models generalize to new datasets, especially if those are trained on large and heterogeneous datasets. Given these findings, measuring brain aging using machine learning methods holds promise for establishing brain-based biomarkers that could aid diagnosis of neurocognitive disorders and be relevant for clinical practice.

## 7. Acknowledgments

The first author thanks all colleagues inside and outside the Max Planck Institute for Human Cognitive and Brain Sciences that provided valuable feedback for this project, especially the members of the Neuroanatomy and Connectivity Group.

We would like to thank the Enhanced Nathan Kline Institute-Rockland Sample initiative for sharing their data.

Franziskus Liem is supported by the Swiss National Science Foundation (SNSF), grant number P2ZHP1_155200. Gaël Varoquaux and Mehdi Rahim are supported by the NiConnect project (ANR-11-BINF-0004 NiConnect). Jana Kynast is supported by the Max-Planck International Research Network on Aging (MaxNetAging).

This work is supported by the European Union, the European Regional Development Fund, and the Free State of Saxony within the framework of the excellence initiative, and LIFE-Leipzig Research Center for Civilization Diseases, University of Leipzig (project numbers 713-241202, 713-241202, 14505/2470, 14575/2470), and by the German Research Foundation (CRC1052 Obesity mechanisms Project A01).

## 8. References

Abraham, A., Milham, M., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. (2016). Deriving robust biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, . doi:10.1016/j.neuroimage.2016.10.045.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. URL: http://www.frontiersin.org/Neuroinformatics/10.3389/fninf.2014.00014/abstract. doi:10.3389/fninf.2014.00014.

Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J., & Raznahan, A. (2016). Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human Brain Mapping*, *37*, 2385–2397. URL: http://doi.wiley.com/10.1002/hbm.23180. doi:10.1002/hbm.23180.

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders . (5th ed.). Washington, DC.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, *54*, 2033–2044. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811910012061. doi:10.1016/j.neuroimage.2010.09.025.

Beck, I. R., Schmid, N. S., Berres, M., & Monsch, A. U. (2014). Establishing robust cognitive dimensions for characterization and differentiation of patients with Alzheimer's disease, mild cognitive impairment, frontotemporal dementia and depression. *International Journal of Geriatric Psychiatry*, *29*, 624–634. URL: http://doi.wiley.com/10.1002/gps.4045. doi:10.1002/gps.4045.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*, 90–101. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811907003837. doi:10.1016/j.neuroimage.2007.04.042.

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, *51*, 1126–1139. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20226257&retmode=ref&cmd=prlinks. doi:10.1016/j.neuroimage.2010.02.082.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289–300. URL: http://www.jstor.org/stable/2346101. doi:10.2307/2346101.

Bölte, S. (2005). Reading the Mind in the Eyes Test Erwachsenenversion - Von Simon Baron-Cohen (2001). Deutsche Bearbeitung.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. URL: http://link.springer.com/10.1023/A:1010933404324. doi:10.1023/A:1010933404324.

Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M. E., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J. R., Garrett, C., Bastos-Leite, A. J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez-Meza, A. M., Dolph, C. V., Iftekharuddin, K. M., Eskildsen, S. F., Coupé, P., Fonov, V. S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K. R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G. M., Stoyanov, Z. V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J. C., Niessen, W. J., Klein, S., & Alzheimer's Disease Neuroimaging Initiative (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. *NeuroImage*, *111*, 562–579. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811915000737. doi:10.1016/j.neuroimage.2015.01.048.

Brown, T. T., Kuperman, J. M., Chung, Y., Erhart, M., McCabe, C., Hagler, D. J., Venkatraman, V. K., Akshoomoff, N., Amaral, D. G., Bloss, C. S., Casey, B. J., Chang, L., Ernst, T. M., Frazier, J. A., Gruen, J. R., Kaufmann, W. E., Kenet, T., Kennedy, D. N., Murray, S. S., Sowell, E. R., Jernigan, T. L., & Dale, A. M. (2012). Neuroanatomical assessment of biological maturity. *Current biology : CB*, *22*, 1693–1698. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22902750&retmode=ref&cmd=prlinks. doi:10.1016/j.cub.2012.07.002.

Bzdok, D. (2016). Classical Statistics and Statistical Learning in Imaging Neuroscience. *arXiv.org*, . URL: http://arxiv.org/abs/1603.01857v1. arXiv:1603.01857v1.

Castellanos, F. X., Di Martino, A., Craddock, R. C., Mehta, A. D., & Milham, M. P. (2013). Clinical applications of the functional connectome. *NeuroImage*, *80*, 527–540. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23631991&retmode=ref&cmd=prlinks. doi:10.1016/j.neuroimage.2013.04.083.

Chiarelli, P. A., Bulte, D. P., Wise, R., Gallichan, D., & Jezzard, P. (2007). A calibration method for quantitative BOLD fMRI based on hyperoxia. *NeuroImage*, *37*, 808–820. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811907004375. doi:10.1016/j.neuroimage.2007.05.033.

Cohen, S. B., Wheelwright, S., & Hill, J. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *... and psychiatry*, . URL: http://onlinelibrary.wiley.com/doi/10.1111/1469-7610.00715/full. doi:10.1111/1469-7610.00715/full.

Cole, J. H., Leech, R., Sharp, D. J., & Alzheimer's Disease Neuroimaging Initiative (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of neurology*, *77*, 571–581. URL: http://doi.wiley.com/10.1002/ana.24367. doi:10.1002/ana.24367.

Craddock, R. C., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *62*, 1619–1628. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19859933&retmode=ref&cmd=prlinks. doi:10.1002/mrm.22159.

Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S., & Milham, M. P. (2013). Imaging human connectomes at the macroscale. *Nature Methods*, *10*, 524–539. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23722212&retmode=ref&cmd=prlinks. doi:10.1038/nmeth.2482.

Damoiseaux, J. S., Beckmann, C. F., Arigita, E. J. S., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., & Rombouts, S. A. R. B. (2008). Reduced resting-state brain activity in the "default network" in normal aging. *Cerebral Cortex*, *18*, 1856–1864. URL: http://www.cercor.oxfordjournals.org/cgi/doi/10.1093/cercor/bhm207. doi:10.1093/cercor/bhm207.

Damoiseaux, J. S., & Greicius, M. D. (2009). Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity. *Brain Structure and Function*, *213*, 525–533. doi:10.1007/s00429-009-0208-6.

Damoiseaux, J. S., Prater, K. E., Miller, B. L., & Greicius, M. D. (2012). Functional connectivity tracks clinical deterioration in Alzheimer's disease. *Neurobiology of Aging*, *33*, 828.e19–30. URL: http://linkinghub.elsevier.com/retrieve/pii/S019745801100251X. doi:10.1016/j.neurobiolaging.2011.06.024.

Dennis, E. L., & Thompson, P. M. (2014). Functional brain connectivity using fMRI in aging and Alzheimer's disease. *Neuropsychology Review*, *24*, 49–62. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24562737&retmode=ref&cmd=prlinks. doi:10.1007/s11065-014-9249-6.

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., Barnes, K. A., Dubis, J. W., Feczko, E., Coalson, R. S., Pruett, J. R., Barch, D. M., Petersen, S. E., & Schlaggar, B. L. (2010). Prediction of individual brain maturity using fMRI. *Science*, *329*, 1358–1361. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1194144. doi:10.1126/science.1194144.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1996). Support Vector Regression Machines. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996* (pp. 155–161). URL: http://papers.nips.cc/paper/1238-support-vector-regression-machines.

Dubois, J., & Adolphs, R. (2016). Building a Science of Individual Differences from fMRI. *Trends in cognitive sciences*, *20*, 425–443. URL: http://linkinghub.elsevier.com/retrieve/pii/S1364661316300079. doi:10.1016/j.tics.2016.03.014.

Erus, G., Battapady, H., Satterthwaite, T. D., Hakonarson, H., Gur, R. E., Davatzikos, C., & Gur, R. C. (2014). Imaging Patterns of Brain Development and their Relationship to Cognition. *Cerebral Cortex*, . URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24421175&retmode=ref&cmd=prlinks. doi:10.1093/cercor/bht425.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*, 774–781. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22248573&retmode=ref&cmd=prlinks. doi:10.1016/j.neuroimage.2012.01.021.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=1202204&retmode=ref&cmd=prlinks.

Fornito, A., Zalesky, A., & Bullmore, E. T. (2010). Network scaling effects in graph analytic studies of human resting-state FMRI data. *Frontiers in Systems Neuroscience*, *4*, 22. URL: http://journal.frontiersin.org/article/10.3389/fnsys.2010.00022/abstract. doi:10.3389/fnsys.2010.00022.

Franke, K., & Gaser, C. (2012). Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease 1. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *25*, 235–245. URL: http://www.psycontent.com/index/79311259G2307246.pdf. doi:10.1024/1662-9647/a000074.

Franke, K., Ristow, M., Gaser, C., & Alzheimer's Disease Neuroimaging Initiative (2014). Gender-specific impact of personal health parameters on individual brain aging in cognitively unimpaired elderly subjects. *Frontiers in Aging Neuroscience*, *6*, 94. URL: http://journal.frontiersin.org/article/10.3389/fnagi.2014.00094/abstract. doi:10.3389/fnagi.2014.00094.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., & Alzheimer's Disease Neuroimaging Initiative (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *NeuroImage*, *50*, 883–892. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20070949&retmode=ref&cmd=prlinks. doi:10.1016/j.neuroimage.2010.01.005.

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., & Turner, R. (1996). Movement-related effects in fMRI time-series. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *35*, 346–355. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=8699946&retmode=ref&cmd=prlinks.

Gabrieli, J. D. E., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, *85*, 11–26. URL: http://linkinghub.elsevier.com/retrieve/pii/S0896627314009672. doi:10.1016/j.neuron.2014.10.047.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., & Alzheimer's Disease Neuroimaging Initiative (2013). BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. *PLoS ONE*, *8*, e67346. URL: http://dx.plos.org/10.1371/journal.pone.0067346. doi:10.1371/journal.pone.0067346.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21897815&retmode=ref&cmd=prlinks. doi:10.3389/fninf.2011.00013.

Greicius, M. (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Current opinion in neurology*, *21*, 424–430. URL: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00019052-200808000-00007. doi:10.1097/WCO.0b013e328306f2c5.

Groves, A. R., Smith, S. M., Fjell, A. M., Tamnes, C. K., Walhovd, K. B., Douaud, G., Woolrich, M. W., & Westlye, L. T. (2012). Benefits of multi-modal fusion analysis on a large-scale dataset: Life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure. *NeuroImage*, *63*, 365–380. URL: http://www.sciencedirect.com/science/article/pii/S1053811912006532. doi:10.1016/j.neuroimage.2012.06.038.

Hogstrom, L. J., Westlye, L. T., Walhovd, K. B., & Fjell, A. M. (2013). The structure of the cerebral cortex across adult life: age-related patterns of surface area, thickness, and gyrification. *Cerebral Cortex*, *23*, 2521–2530. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22892423&retmode=ref&cmd=prlinks. doi:10.1093/cercor/bhs231.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*, 782–790. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21979382&retmode=ref&cmd=prlinks. doi:10.1016/j.neuroimage.2011.09.015.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., Pantelis, C., & Meisenzahl, E. (2013). Accelerated Brain Aging in Schizophrenia and Beyond: A Neuroanatomical Marker of Psychiatric Disorders. *Schizophrenia bulletin*, *40*, sbt142–1153. URL: http://schizophreniabulletin.oxfordjournals.org/content/early/2013/10/11/schbul.sbt142.full. doi:10.1093/schbul/sbt142.

Liem, F., Mérillat, S., Bezzola, L., Hirsiger, S., Philipp, M., Madhyastha, T., & Jäncke, L. (2015). Reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly. *NeuroImage*, *108*, 95–109. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811914010271. doi:10.1016/j.neuroimage.2014.12.035.

Loeffler, M., Engel, C., Ahnert, P., Alfermann, D., Arélin, K., Baber, R., Beutner, F., Binder, H., Brähler, E., Burkhardt, R., Ceglarek, U., Enzenbach, C., Fuchs, M., Glaesmer, H., Girlich, F., Hagendorff, A., Häntzsch, M., Hegerl, U., Henger, S., Hensch, T., Hinz, A., Holzendorf, V., Husser, D., Kersting, A., Kiel, A., Kirsten, T., Kratzsch, J., Krohn, K., Luck, T., Melzer, S., Netto, J., Nüchter, M., Raschpichler, M., Rauscher, F. G., Riedel-Heller, S. G., Sander, C., Scholz, M., Schönknecht, P., Schroeter, M. L., Simon, J.-C., Speer, R., Stäker, J., Stein, R., Stöbel-Richter, Y., Stumvoll, M., Tarnok, A., Teren, A., Teupser, D., Then, F. S., Tönjes, A., Treudler, R., Villringer, A., Weissgerber, A., Wiedemann, P., Zachariae, S., Wirkner, K., & Thiery, J. (2015). The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. *BMC public health*, *15*, 691. URL: http://www.biomedcentral.com/1471-2458/15/691. doi:10.1186/s12889-015-1983-z.

Luders, E., Cherbuin, N., & Gaser, C. (2016). Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners. *NeuroImage*, *134*, 508–513. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811916300404. doi:10.1016/j.neuroimage.2016.04.007.

Lutti, A., Dick, F., Sereno, M. I., & Weiskopf, N. (2014). Using high-resolution quantitative mapping of R1 as an index of cortical myelination. *NeuroImage*, *93 Pt 2*, 176–188. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811913006423. doi:10.1016/j.neuroimage.2013.06.005.

Meyer, M., Liem, F., Hirsiger, S., Jäncke, L., & Hänggi, J. (2014). Cortical surface area and cortical thickness demonstrate differential structural asymmetry in auditory-related areas of the human cortex. *Cerebral Cortex*, *24*, 2541–2552. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23645712&retmode=ref&cmd=prlinks. doi:10.1093/cercor/bht094.

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., Mellits, E. D., & Clark, C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assesment of Alzheimer's disease. *Neurology*, *39*, 1159–1159. URL: http://www.neurology.org/cgi/doi/10.1212/WNL.39.9.1159. doi:10.1212/WNL.39.9.1159.

Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R. T., Kamiel, S. M., Anwar, A. R., Hinz, C. M., Kaplan, M. S., Rachlin, A. B., Adelsberg, S., Cheung, B., Khanuja, R., Yan, C., Craddock, C. C., Calhoun, V., Courtney, W., King, M., Wood, D., Cox, C. L., Kelly, A. M. C., Di Martino, A., Petkova, E., Reiss, P. T., Duan, N., Thomsen, D., Biswal, B., Coffey, B., Hoptman, M. J., Javitt, D. C., Pomara, N., Sidtis, J. J., Koplewicz, H. S., Castellanos, F. X., Leventhal, B. L., & Milham, M. P. (2012). The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Frontiers in Neuroscience*, *6*, 152. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23087608&retmode=ref&cmd=prlinks. doi:10.3389/fnins.2012.00152.

Pedregosa, F., Varoquaux, G., & Gramfort, A. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning*, *12*, 2825–2830. URL: http://dl.acm.org/citation.cfm?id=2078195.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*, S199–209. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19070668&retmode=ref&cmd=prlinks. doi:10.1016/j.neuroimage.2008.11.007.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*, 2142–2154. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811911011815. doi:10.1016/j.neuroimage.2011.10.018.

Rahim, M., Thirion, B., Comtat, C., & Varoquaux, G. (2016). Transmodal Learning of Functional Networks for Alzheimer's Disease Prediction. *IEEE Journal on Selected Topics in Signal Processing*, . URL: https://hal.inria.fr/hal-01353728.

Raz, N., & Rodrigue, K. M. (2006). Differential aging of the brain: patterns, cognitive correlates and modifiers. *Neuroscience and biobehavioral reviews*, *30*, 730–748. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=16919333&retmode=ref&cmd=prlinks. doi:10.1016/j.neubiorev.2006.07.001.

Reitan, R. (1979). *Trail-making test*. Arizona: Reitan Neuropsychology Labo-

ratory.

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. W., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, *107*, 107–115. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811914009975. doi:10.1016/j.neuroimage.2014.12.006.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, *64*, 240–256. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811912008609. doi:10.1016/j.neuroimage.2012.08.052.

Schroeter, M. L., Zysset, S., Kupka, T., Kruggel, F., & Yves von Cramon, D. (2002). Near-infrared spectroscopy can detect brain activity during a color-word matching Stroop task in an event-related design. *Human Brain Mapping*, *17*, 61–71. URL: http://doi.wiley.com/10.1002/hbm.10052. doi:10.1002/hbm.10052.

Shehzad, Z., Kelly, A. M. C., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., Lee, S. H., Margulies, D. S., Roy, A. K., Biswal, B. B., Petkova, E., Castellanos, F. X., & Milham, M. P. (2009). The resting brain: unconstrained yet reliable. *Cerebral Cortex*, *19*, 2209–2229. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19221144&retmode=ref&cmd=prlinks. doi:10.1093/cercor/bhn256.

Steffener, J., Habeck, C., O'Shea, D., Razlighi, Q., Bherer, L., & Stern, Y. (2016). Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiology of Aging*, *40*, 138–144. URL: http://linkinghub.elsevier.com/retrieve/pii/S0197458016000233. doi:10.1016/j.neurobiolaging.2016.01.014.

Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., & Walhovd, K. B. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change. *Journal of Neuroscience*, *34*, 8488–8498. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0391-14.2014. doi:10.1523/JNEUROSCI.0391-14.2014.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. URL: http://content.apa.org/journals/xge/18/6/643. doi:10.1037/h0054651.

Thalmann, B., Monsch, A. U., Bernasconi, F., Berres, M., Schneitter, M., Ermini-Fünfschilling, D., Spiegel, R., & Stähelin, H. B. (1997). *Die CERAD Neuropsychologische Testbatterie – Ein gemeinsames minimales Instrumentarium zur Demenzabklärung*. Basel: Memory Clinic, Geriatrische Universitätsklinik.

Thirion, B., Varoquaux, G., Dohmatob, E., & Poline, J.-B. (2014). Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience*, *8*, 167. URL: http://journal.frontiersin.org/article/10.3389/fnins.2014.00167/abstract. doi:10.3389/fnins.2014.00167.

Treisman, A., & Fearnley, S. (1969). The Stroop test: selective attention to colours and words. *Nature*, *222*, 437–439. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=5768618&retmode=ref&cmd=prlinks.

Ullman, H., Almeida, R., & Klingberg, T. (2014). Structural Maturation and Brain Activity Predict Future Working Memory Capacity during Childhood Development. *Journal of Neuroscience*, *34*, 1592–1598. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24478343&retmode=ref&cmd=prlinks. doi:10.1523/JNEUROSCI.0842-13.2014.

Van Dijk, K. R. A., Sabuncu, M. R., & Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage*, *59*, 431–438. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21810475&retmode=ref&cmd=prlinks. doi:10.1016/j.neuroimage.2011.07.044.

Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience*, *3*, 28. URL: http://www.gigasciencejournal.com/content/3/1/28. doi:10.1186/2047-217X-3-28.

Waskom, M., Botvinnik, O., drewokane, Hobson, P., Halchenko, Y., Lukauskas, S., Warmenhoven, J., Cole, J. B., Hoyer, S., Vanderplas, J., gkunter, Villalba, S., Quintero, E., Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Evans, C., Fitzgerald, C., Nagy, T., Ziegler, E., Megies, T., Wehner, D., St-Jean, S., Coelho, L. P., Hitz, G., Lee, A., & Rocher, L. (2016). *seaborn: v0.7.0 (January 2016)*. Technical Report. URL: http://dx.doi.org/10.5281/zenodo.45133. doi:10.5281/zenodo.45133.

Yarkoni, T., & Westfall, J. (2016). Choosing prediction over explanation in psychology: Lessons from machine learning, . URL: http://jakewestfall.org/publications/Yarkoni_Westfall_choosing_prediction.pdf. doi:10.1242/dmm.006627.

Ziegler, G., Dahnke, R., Jäncke, L., Yotter, R. A., May, A., & Gaser, C. (2012). Brain structural trajectories over the adult lifespan. *Human Brain Mapping*, *33*, 2377–2389. URL: http://onlinelibrary.wiley.com/doi/10.1002/hbm.21374/full. doi:10.1002/hbm.21374.

Zysset, S., Müller, K., Lohmann, G., & von Cramon, D. Y. (2001). Color-word matching stroop task: separating interference and response conflict. *NeuroImage*, *13*, 29–36. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811900906657. doi:10.1006/nimg.2000.0665.

## Appendix A. Supplementary methods

### *Appendix A.1. Tuning curve*

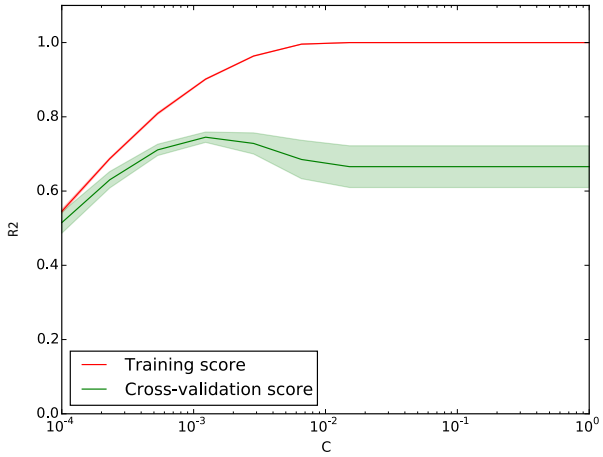


Figure A.6: Exemplarily, a tuning curve is presented for *cortical thickness*. The tuning curve was run on the training set. The support vector regression's $C$ parameter (x-axis) was varied, the R2 (coefficient of determination, y-axis) was evaluated as training and cross-validation score. A 'sweet spot', a maximum cross-validation score (green line) only a slightly better training performance (red line), can be seen around $C = 10^{-3}$.

### *Appendix A.2. NKI sample*

Data from the enhanced Nathan Kline Institute - Rockland sample (`fcon_1000.projects.nitrc.org/indi/enhanced`, Nooner et al., 2012) was used in parts of this study. Subjects selected for the present work (N = 475) were between 18 and 85 years (M = 45.78; SD = 18.91; 311 female, 161 male).

### *Appendix A.3. MR data*

Brain imaging was performed on a a 3T Siemens Trio scanner with a 32 channel head coil.

T2*-weighted functional images were acquired using an multiband echo-planar-imaging sequence with 3 mm isotropic voxels, 40 slices, echo time (TE) of 30 ms, repetition time (TR) of 645 ms, multiband acceleration factor of 4 and a flip-angle of 60° (`fcon_1000.projects.nitrc.org/indi/enhanced/mri_protocol.html`). The resting-state sequence lasted approximately 9.5 min (900 volumes), during which subjects were instructed to keep their eyes open and not to fall asleep. No fieldmaps have been acquired.

High resolution T1-weighted structural images were acquired using the MP-RAGE sequence with 1 mm isotropic voxels, 176 slices, a TR of 1900 ms, and a TE of 2.52 ms.

### *Appendix A.4. MR data preprocessing*

Preprocessing was performed very similar to preprocessing of the LIFE sample detailed in section 2.4. Since no fieldmaps were available for NKI and data from the two studies have been preprocessed independently, minor differences exist: CompCor was performed with five components instead of six and normalization into standard space was performed with FSL's FNIRT, not ANTS. Preprocessing scripts are available at `github.com/fliem/nki_nilearn`.

## Appendix B. Supplementary results

### *Appendix B.1. Multimodal data increases age prediction performance*

| | M(APE) | SD(APE) | R2 | W | p(FDR) |
|---|---|---|---|---|---|
| connectivity matrix 197 | 5.99 | 4.57 | 0.75 | 2e+05 | 8.2e-38 |
| connectivity matrix 444 | 5.77 | 4.42 | 0.77 | 2.1e+05 | 9.1e-31 |
| stacked-function | 5.25 | 4.40 | 0.80 | 2.3e+05 | 1.6e-22 |
| | | | | | |
| cortical thickness | 5.95 | 4.69 | 0.75 | 2.1e+05 | 9.9e-32 |
| cortical surface area | 7.29 | 5.95 | 0.62 | 1.7e+05 | 5.9e-53 |
| subcortical | 6.44 | 5.02 | 0.71 | 1.9e+05 | 1.6e-40 |
| stacked-anatomy | 4.83 | 4.01 | 0.83 | 2.7e+05 | 1.8e-10 |
| | | | | | |
| stacked-multimodal | 4.29 | 3.49 | 0.87 | - | - |

Table B.1: Age prediction (absolute prediction error (APE)) on the test sample. Statistical test against best model (Wilcoxon signed-rank test against stacked-multimodal; N = 1177). See Figure 2.

### *Appendix B.2. RF feature importance*

| | stacked-function | stacked-anatomy | stacked-multimodal |
|---|---|---|---|
| connectivity matrix 197 | 0.12 | - | 0.09 |
| connectivity matrix 444 | 0.88 | - | 0.68 |
| cortical thickness | - | 0.53 | 0.14 |
| cortical surface area | - | 0.03 | 0.01 |
| subcortical | - | 0.44 | 0.08 |

Table B.2: Values of feature importance from the Random Forest multi-source models showing the contribution of the single-source data.

### *Appendix B.3. Cross-validation (CV) scores*

| | M(APE) | SD(APE) | M(R2) | SD(R2) |
|---|---|---|---|---|
| connectivity matrix 197 | 6.33 | 0.31 | 0.72 | 0.03 |
| connectivity matrix 444 | 6.02 | 0.40 | 0.75 | 0.03 |
| stacked-function | 5.45 | 0.20 | 0.78 | 0.02 |
| | | | | |
| cortical thickness | 6.25 | 0.12 | 0.73 | 0.02 |
| cortical surface area | 7.21 | 0.34 | 0.61 | 0.04 |
| subcortical | 6.60 | 0.34 | 0.69 | 0.04 |
| stacked-anatomy | 4.98 | 0.10 | 0.82 | 0.02 |
| | | | | |
| stacked-multimodal | 4.53 | 0.12 | 0.85 | 0.01 |

Table B.3: CV scores of five folds run on training sample.

### *Appendix B.4. Increased brain aging in impaired subjects*

| OCI group | N | M(Age) | SD(Age) | Sex(f/m) |
|---|---|---|---|---|
| norm | 729 | 59.2 | 15.2 | 364/365 |
| mild | 632 | 58.0 | 14.9 | 294/338 |
| major | 251 | 58.3 | 15.7 | 115/136 |

Table B.4: Sample characteristics of OCI (objective cognitive impairment) groups of the test sample.

| | OCI norm | | OCI mild | | OCI major | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | M(APE) | SD(APE) | M(APE) | SD(APE) | H | p(FDR) |
| connectivity matrix 197 | 6.61 | 5.28 | 6.22 | 4.86 | 7.09 | 5.91 | 1.9 | 0.61 |
| connectivity matrix 444 | 6.21 | 4.95 | 5.87 | 4.64 | 6.83 | 5.53 | 4.1 | 0.35 |
| stacked-function | 5.69 | 4.73 | 5.54 | 4.74 | 6.20 | 5.64 | 0.95 | 0.62 |
| cortical thickness | 5.87 | 4.77 | 6.17 | 5.18 | 7.11 | 5.80 | 6.8 | 0.13 |
| cortical surface area | 7.62 | 6.41 | 7.58 | 6.66 | 8.34 | 7.33 | 1 | 0.62 |
| subcortical | 6.16 | 4.74 | 6.47 | 5.19 | 7.73 | 6.50 | 7.1 | 0.13 |
| stacked-anatomy | 4.76 | 3.91 | 4.90 | 4.08 | 5.43 | 5.23 | 1.1 | 0.62 |
| stacked-multimodal | 4.40 | 3.60 | 4.48 | 3.72 | 5.16 | 4.81 | 2 | 0.61 |

Table B.5: Differences in absolute prediction error (APE) between objective cognitive impairment (OCI) groups (Kruskal-Wallis H-test: effect of OCI group (norm, mild, major) on absolute prediction error (APE); N(norm) = 729, N(mild) = 632, N(major) = 251; N(training) = 724.

| | OCI norm | | OCI mild | | OCI major | | | |
|---|---|---|---|---|---|---|---|---|
| | M(BA) | SD(BA) | M(BA) | SD(BA) | M(BA) | SD(BA) | H | p(FDR) |
| connectivity matrix 197 | 0.48 | 8.45 | 1.31 | 7.79 | 2.22 | 8.97 | 7.9 | 0.026 |
| connectivity matrix 444 | -0.04 | 7.95 | 0.75 | 7.45 | 1.52 | 8.66 | 7.5 | 0.027 |
| stacked-function | -0.24 | 7.40 | 0.64 | 7.26 | 1.22 | 8.30 | 5.1 | 0.077 |
| cortical thickness | 0.73 | 7.54 | 2.16 | 7.76 | 2.99 | 8.68 | 21 | 5.1e-05 |
| cortical surface area | 1.32 | 9.87 | 2.82 | 9.69 | 3.31 | 10.61 | 15 | 0.001 |
| subcortical | -0.12 | 7.78 | 1.39 | 8.18 | 3.91 | 9.32 | 44 | 2.1e-09 |
| stacked-anatomy | -0.52 | 6.14 | 0.76 | 6.33 | 1.88 | 7.31 | 26 | 7.8e-06 |
| stacked-multimodal | -0.38 | 5.68 | 0.74 | 5.78 | 1.72 | 6.85 | 22 | 4e-05 |

Table B.6: Differences in brain aging (BA) between objective cognitive impairment (OCI) groups (Kruskal-Wallis H-test: effect of OCI group (norm, mild, major) on brain age; N(norm) = 729, N(mild) = 632, N(major) = 251). N(training) = 724. See Figure B.7.

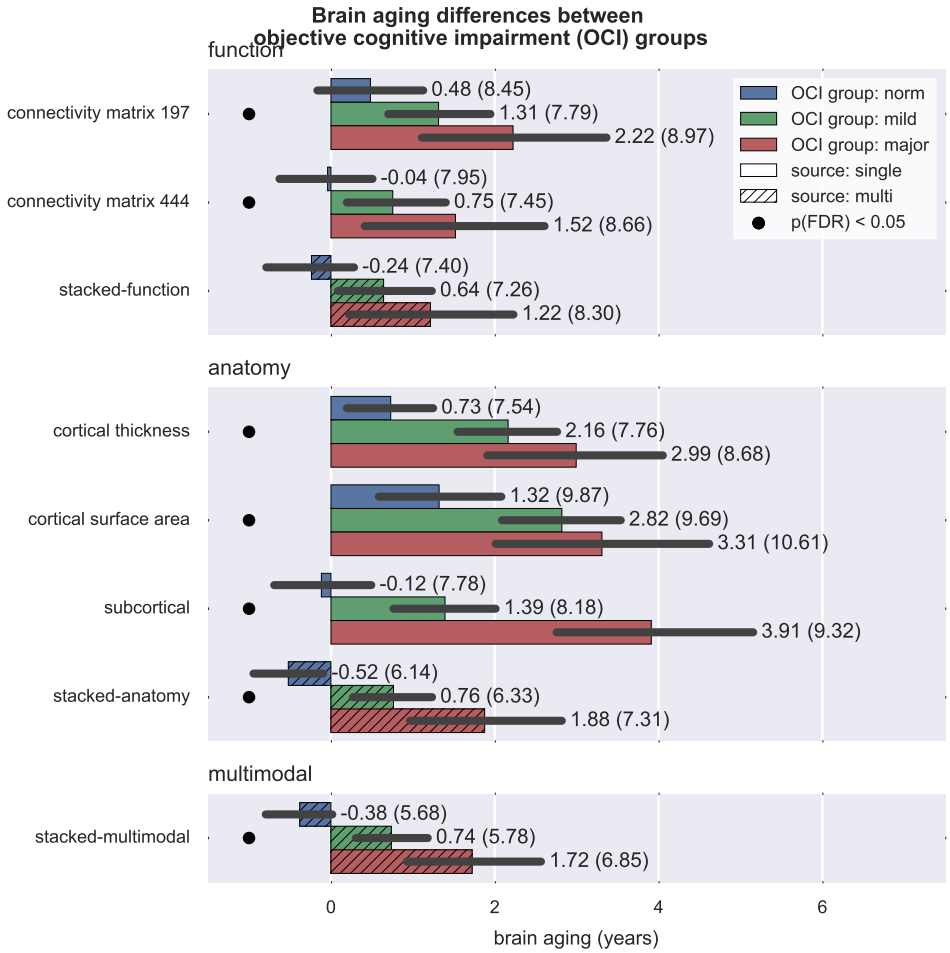


Figure B.7: Differences in brain aging between objective cognitive impairment (OCI) groups. Positive brain aging values indicate that a brain appears older than expected from chronological age. Note that for the majority of models impairment measured by OCI is significantly related to higher brain aging (as indicated by the black dot), i.e., more advanced brain aging. For full statistics see Table B.6.

### *Appendix B.5. Robustness against head motion*

#### *Appendix B.5.1. Motion regression*

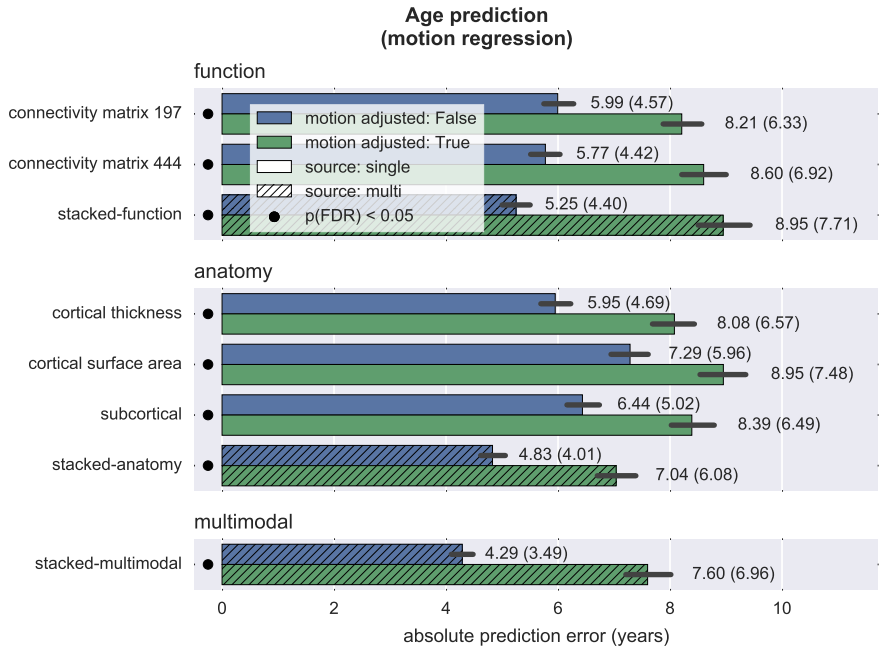


Figure B.8: Motion adjustment via motion regression. Absolute prediction error significantly increases if motion is regressed out of brain data (black dots). Motion adjusted: True: analysis includes motion regression on brain data; False: original analysis without motion regression.

| | motion adjusted False | | | motion adjusted True | | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | R2 | M(APE) | SD(APE) | R2 | W | p(FDR) |
| connectivity matrix 197 | 5.99 | 4.57 | 0.75 | 8.21 | 6.33 | 0.53 | 1.9e+05 | 1.3e-39 |
| connectivity matrix 444 | 5.77 | 4.42 | 0.77 | 8.60 | 6.92 | 0.47 | 1.8e+05 | 1.7e-46 |
| stacked-function | 5.25 | 4.40 | 0.80 | 8.95 | 7.71 | 0.40 | 1.7e+05 | 3.2e-48 |
| cortical thickness | 5.95 | 4.69 | 0.75 | 8.08 | 6.57 | 0.53 | 1.9e+05 | 1e-42 |
| cortical surface area | 7.29 | 5.95 | 0.62 | 8.95 | 7.48 | 0.41 | 2.1e+05 | 3.6e-33 |
| subcortical | 6.44 | 5.02 | 0.71 | 8.39 | 6.48 | 0.51 | 2.1e+05 | 2e-33 |
| stacked-anatomy | 4.83 | 4.01 | 0.83 | 7.04 | 6.08 | 0.63 | 2.1e+05 | 3.6e-33 |
| stacked-multimodal | 4.29 | 3.49 | 0.87 | 7.60 | 6.95 | 0.54 | 1.8e+05 | 7.3e-47 |

Table B.7: Motion adjustment via motion regression. Absolute prediction error (APE) significantly increases if motion is regressed out of brain data. Motion adjusted: True: analysis includes motion regression on brain data; False: original analysis without motion regression. R2: coefficient of determination. (Wilcoxon signed-rank test: motion adjusted False against True; N = 1177). See Figure B.8.

#### *Appendix B.5.2. Motion matching*

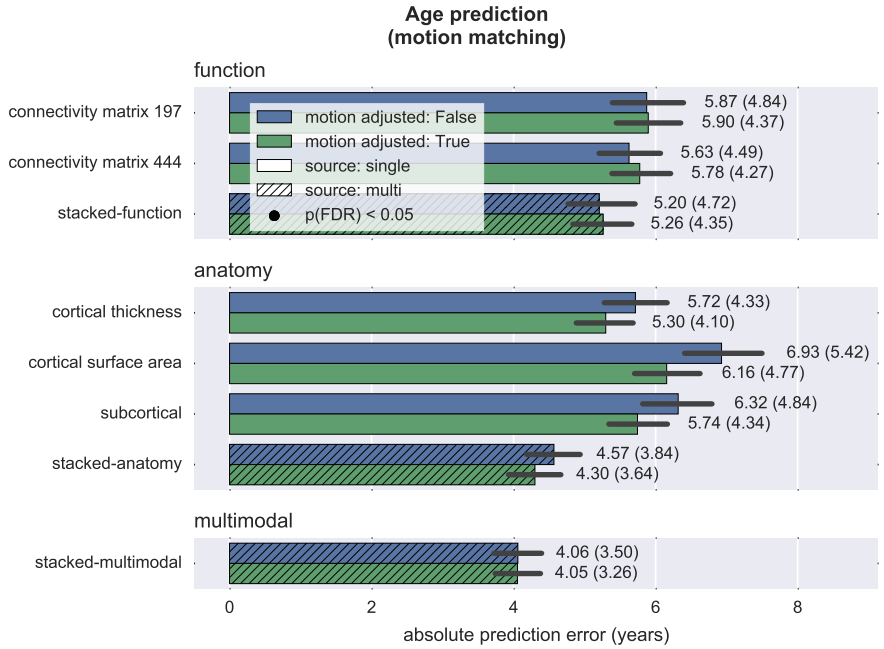


Figure B.9: Motion adjustment via motion matching. Motion adjusted: True: sample without *age* × *motion* correlation; False: sample with preserved *age* × *motion* correlation. Note that for all models prediction with and without motion matching is equally good, indicating that the models' predictions are not driven by head motion. For full statistics see Table B.8.

| | motion adjusted False | | | motion adjusted True | | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | R2 | M(APE) | SD(APE) | R2 | U | p(FDR) |
| connectivity matrix 197 | 5.87 | 4.83 | 0.70 | 5.90 | 4.36 | 0.61 | 7.3e+04 | 0.65 |
| connectivity matrix 444 | 5.63 | 4.48 | 0.73 | 5.78 | 4.26 | 0.63 | 7.3e+04 | 0.65 |
| stacked-function | 5.20 | 4.71 | 0.74 | 5.26 | 4.35 | 0.66 | 7.3e+04 | 0.65 |
| cortical thickness | 5.72 | 4.32 | 0.73 | 5.30 | 4.10 | 0.67 | 7.9e+04 | 0.49 |
| cortical surface area | 6.93 | 5.41 | 0.60 | 6.16 | 4.76 | 0.56 | 8e+04 | 0.49 |
| subcortical | 6.32 | 4.83 | 0.67 | 5.74 | 4.34 | 0.62 | 7.9e+04 | 0.49 |
| stacked-anatomy | 4.57 | 3.83 | 0.82 | 4.30 | 3.63 | 0.77 | 7.8e+04 | 0.65 |
| stacked-multimodal | 4.06 | 3.50 | 0.85 | 4.05 | 3.26 | 0.80 | 7.4e+04 | 0.86 |

Table B.8: Motion adjustment via motion matching. Note that for all models prediction with and without motion matching is equally good, indicating that the models' predictions are not driven by head motion. Motion adjusted: True: sample without *age* × *motion* correlation; False: sample with preserved *age* × *motion* correlation. (Mann-Whitney U test: motion adjusted False against True; N(False) = 387, N(True) = 387). See Figure B.9.

*Appendix B.6. Generalization to new site*

*Appendix B.6.1. Generalization to new site: One sample training*

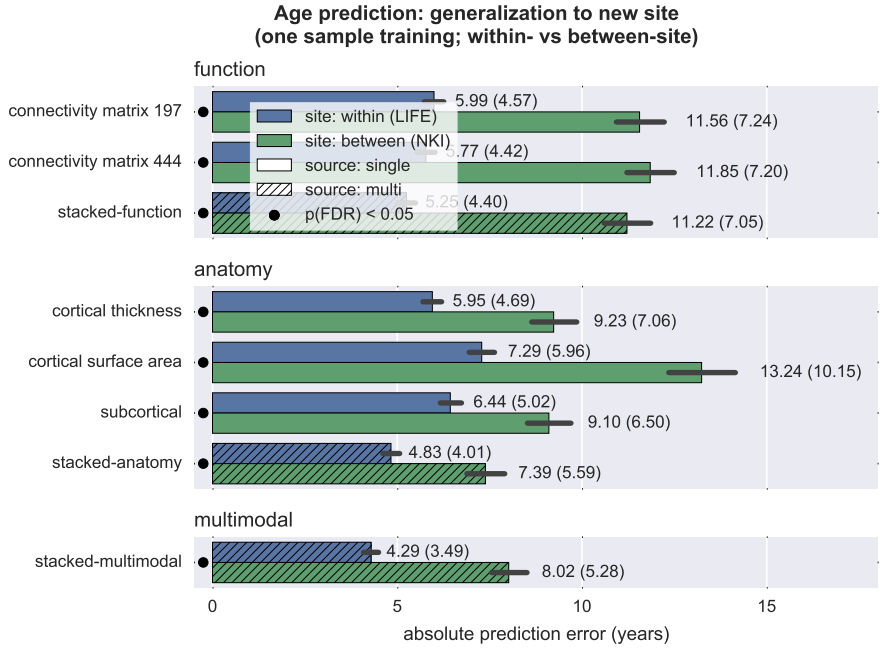


Figure B.10: Standard training procedure (one sample training) showed significantly ($p(FDR) < 0.05$ as indicated by the black dot) better prediction performance in LIFE data (within site) than in NKI data (between site). See Table B.9).

| | site within (LIFE) | | | site between (NKI) | | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | R2 | M(APE) | SD(APE) | R2 | U | p(FDR) |
| connectivity matrix 197 | 5.99 | 4.57 | 0.75 | 11.56 | 7.23 | 0.48 | 1.5e+05 | 7.4e-49 |
| connectivity matrix 444 | 5.77 | 4.42 | 0.77 | 11.85 | 7.19 | 0.46 | 1.4e+05 | 1.7e-57 |
| stacked-function | 5.25 | 4.40 | 0.80 | 11.22 | 7.04 | 0.51 | 1.3e+05 | 5.4e-63 |
| cortical thickness | 5.95 | 4.69 | 0.75 | 9.23 | 7.06 | 0.62 | 2.1e+05 | 4.4e-17 |
| cortical surface area | 7.29 | 5.95 | 0.62 | 13.24 | 10.14 | 0.22 | 1.9e+05 | 8.3e-26 |
| subcortical | 6.44 | 5.02 | 0.71 | 9.10 | 6.49 | 0.65 | 2.1e+05 | 1.4e-14 |
| stacked-anatomy | 4.83 | 4.01 | 0.83 | 7.39 | 5.59 | 0.76 | 2e+05 | 7.5e-19 |
| stacked-multimodal | 4.29 | 3.49 | 0.87 | 8.02 | 5.27 | 0.74 | 1.6e+05 | 2.7e-43 |

Table B.9: Generalization to new site. Standard training procedure (one sample training) showed significantly better prediction performance in LIFE data (within site) than in NKI data (between site) (Mann-Whitney U test: site within (LIFE) against between (NKI); N(within (LIFE)) = 1177, N(between (NKI)) = 475). See Figure B.10.

*Appendix B.6.2. Generalization to new site: Two samples training*

| | training one sample | | | training two samples | | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | R2 | M(APE) | SD(APE) | R2 | W | p(FDR) |
| connectivity matrix 197 | 11.53 | 7.30 | 0.48 | 10.35 | 7.15 | 0.56 | 3.3e+04 | 1.3e-06 |
| connectivity matrix 444 | 11.74 | 7.22 | 0.47 | 10.17 | 6.94 | 0.57 | 2.2e+04 | 9e-20 |
| stacked-function | 11.12 | 7.05 | 0.51 | 9.90 | 6.88 | 0.59 | 3.8e+04 | 0.0015 |
| cortical thickness | 9.21 | 7.03 | 0.62 | 8.11 | 5.93 | 0.72 | 3.3e+04 | 1.3e-06 |
| cortical surface area | 13.23 | 10.20 | 0.22 | 11.64 | 8.29 | 0.43 | 3.2e+04 | 3.5e-07 |
| subcortical | 9.04 | 6.50 | 0.65 | 8.01 | 6.18 | 0.71 | 3.8e+04 | 0.0012 |
| stacked-anatomy | 7.37 | 5.50 | 0.76 | 6.88 | 5.25 | 0.79 | 4.3e+04 | 0.17 |
| stacked-multimodal | 8.00 | 5.21 | 0.74 | 6.93 | 5.08 | 0.79 | 3.6e+04 | 0.00012 |

Table B.10: Generalization to new site. Comparing test prediction performance on NKI data (between site); training on one vs two sites (Wilcoxon signed-rank test: training one sample against two samples; N = 429).

| | site within (LIFE) | | | site between (NKI) | | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | R2 | M(APE) | SD(APE) | R2 | U | p(FDR) |
| connectivity matrix 197 | 6.51 | 5.03 | 0.71 | 10.35 | 7.15 | 0.56 | 1.7e+05 | 1.5e-21 |
| connectivity matrix 444 | 5.87 | 4.53 | 0.76 | 10.17 | 6.94 | 0.57 | 1.6e+05 | 5.8e-29 |
| stacked-function | 5.32 | 4.49 | 0.79 | 9.90 | 6.88 | 0.59 | 1.5e+05 | 1.7e-37 |
| cortical thickness | 6.27 | 4.88 | 0.73 | 8.11 | 5.93 | 0.72 | 2.1e+05 | 7.3e-08 |
| cortical surface area | 7.81 | 6.13 | 0.57 | 11.64 | 8.29 | 0.43 | 1.8e+05 | 1e-16 |
| subcortical | 7.91 | 6.50 | 0.55 | 8.01 | 6.18 | 0.71 | 2.5e+05 | 0.51 |
| stacked-anatomy | 5.11 | 4.27 | 0.81 | 6.88 | 5.25 | 0.79 | 2e+05 | 2.2e-09 |
| stacked-multimodal | 4.46 | 3.70 | 0.85 | 6.93 | 5.08 | 0.79 | 1.8e+05 | 1.8e-19 |

Table B.11: Generalization to new site. After training the model on a mixed-site sample (two sample training, $N_{training,LIFE} = 1177$; $N_{training,NKI} = 46$), predictions on the NKI data improve, but the predictions on the main training site LIFE (within site) still are significantly better than on the minor training site NKI (between site). (Mann-Whitney U test: site within (LIFE) against between (NKI); N(within (LIFE)) = 1177, N(between (NKI)) = 429). See Figure B.11.

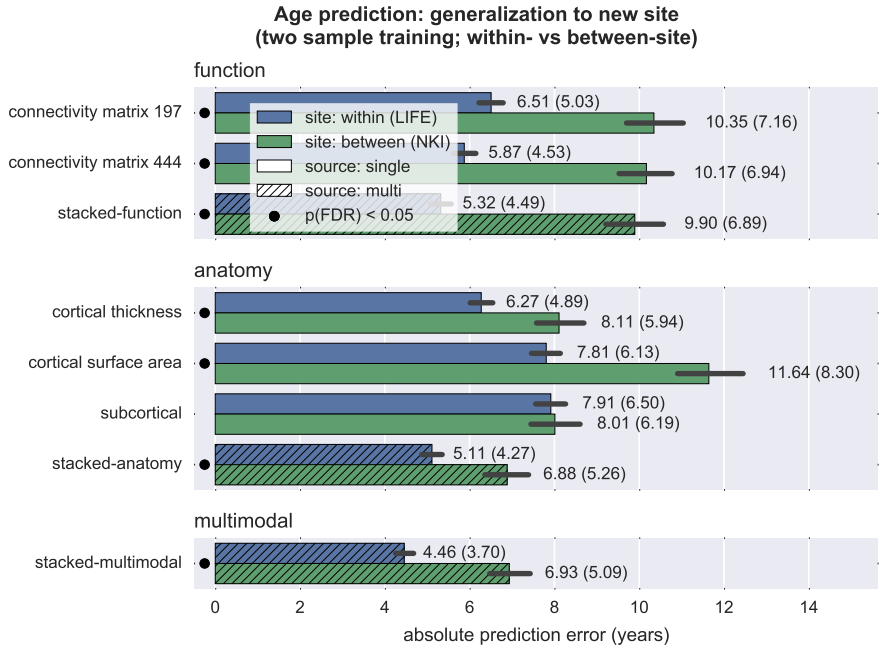


Figure B.11: After training the model on a mixed-site sample (two sample training, $N_{training,LIFE} = 1177$; $N_{training,NKI} = 46$), predictions on the NKI data improve (Table B.10), but the predictions on the main training site LIFE (within site) still are significantly better than on the minor training site NKI (between site).

*Appendix B.6.3. Generalization to new site: Training on full LIFE sample*

| | training one sample | | | training two samples | | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | R2 | M(APE) | SD(APE) | R2 | W | p(FDR) |
| connectivity matrix 197 | 11.31 | 7.12 | 0.50 | 9.75 | 6.74 | 0.60 | 2.9e+04 | 1.2e-10 |
| connectivity matrix 444 | 11.05 | 7.17 | 0.51 | 9.45 | 6.50 | 0.63 | 2.5e+04 | 3.7e-16 |
| stacked-function | 10.27 | 6.59 | 0.58 | 8.88 | 6.41 | 0.66 | 3.7e+04 | 0.00028 |
| cortical thickness | 8.59 | 6.51 | 0.67 | 7.79 | 5.93 | 0.73 | 3.3e+04 | 4.2e-07 |
| cortical surface area | 12.08 | 9.24 | 0.35 | 11.08 | 7.89 | 0.48 | 3.5e+04 | 5.4e-05 |
| subcortical | 8.97 | 6.50 | 0.65 | 8.11 | 6.23 | 0.71 | 3.9e+04 | 0.0046 |
| stacked-anatomy | 7.35 | 5.29 | 0.77 | 6.74 | 4.94 | 0.80 | 4.2e+04 | 0.1 |
| stacked-multimodal | 7.79 | 4.96 | 0.76 | 6.56 | 4.49 | 0.82 | 3.5e+04 | 5.4e-05 |

Table B.12: Generalization to new site; training on full LIFE sample ($N_{training,LIFE}$ = 2377). Comparing test prediction performance on NKI data (between site); training on one sample (LIFE sample only) vs two samples (LIFE + NKI samples; $N_{training,NKI}$ = 46. (Wilcoxon signed-rank test: training one sample against two samples; N = 429). See Figure B.12.

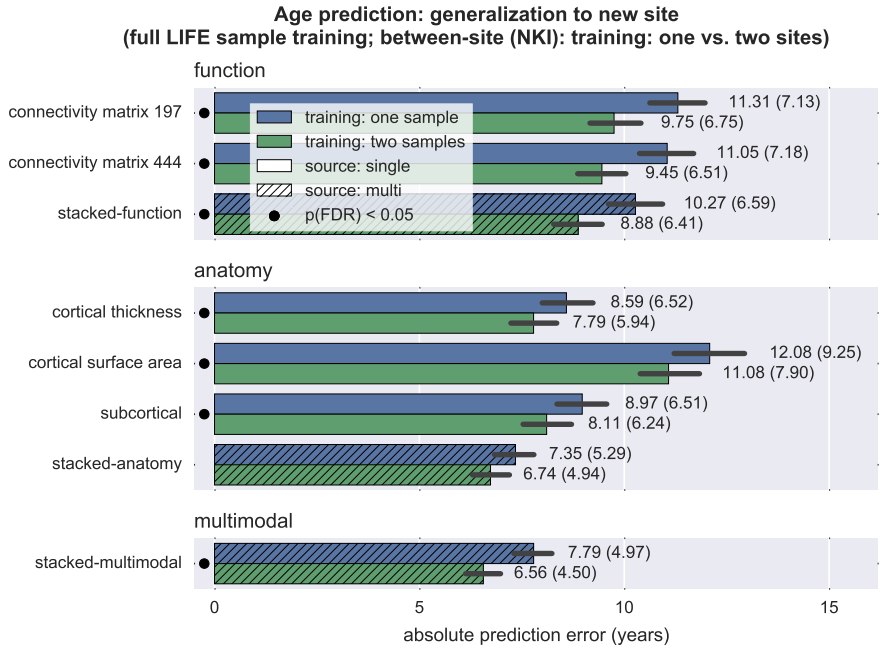


Figure B.12: Generalization to new site; training on full LIFE sample ($N_{training,LIFE}$ = 2377). Comparing test prediction performance on NKI data (between site); training on one sample (LIFE sample only) vs two samples (LIFE + NKI samples; $N_{training,NKI}$ = 46. See Table B.12.

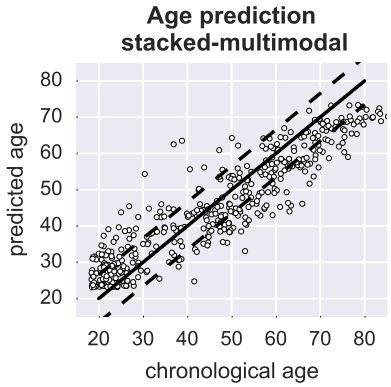


Figure B.13: Chronological and predicted age for the NKI test set from the *stacked-multimodal* model with the two sample training approach. Circles represent subjects, the solid line the perfect prediction, dashed lines the mean absolute prediction error (6.56 years).

### *Appendix B.7. Robustness of two sample training approach*

| | within (LIFE) | | | | between (NKI) | | | |
|---|---|---|---|---|---|---|---|---|
| | M(APE) | SD(APE) | M(R2) | SD(R2) | M(APE) | SD(APE) | M(R2) | SD(R2) |
| connectivity matrix 197 | 6.33 | 0.09 | 0.72 | 0.01 | 10.32 | 0.23 | 0.57 | 0.01 |
| connectivity matrix 444 | 5.85 | 0.10 | 0.76 | 0.01 | 10.58 | 0.27 | 0.55 | 0.02 |
| stacked-function | 5.29 | 0.08 | 0.79 | 0.01 | 9.94 | 0.20 | 0.59 | 0.02 |
| cortical thickness | 6.39 | 0.12 | 0.72 | 0.01 | 8.20 | 0.17 | 0.71 | 0.01 |
| cortical surface area | 7.62 | 0.13 | 0.59 | 0.01 | 11.21 | 0.20 | 0.47 | 0.01 |
| subcortical | 7.65 | 0.25 | 0.58 | 0.04 | 8.58 | 0.31 | 0.67 | 0.04 |
| stacked-anatomy | 5.24 | 0.12 | 0.80 | 0.01 | 7.51 | 0.39 | 0.76 | 0.02 |
| stacked-multimodal | 4.57 | 0.12 | 0.85 | 0.01 | 7.26 | 0.37 | 0.78 | 0.02 |

Table B.13: Stability of two sample training (cf. Table B.11) over ten random splits.