NeuroImage 270 (2023) 119947

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Brain-age prediction: A systematic comparison of machine learning workflows ☆

Shammi More [a,b], Georgios Antonopoulos [a,b], Felix Hoffstaedter [a,b], Julian Caspers [c], Simon B. Eickhoff [a,b], Kaustubh R. Patil [a,b,*], for the Alzheimer's Disease Neuroimaging Initiative

[a] Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany
[b] Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[c] Department of Diagnostic and Interventional Radiology, University Hospital Düsseldorf, Düsseldorf, Germany

ARTICLE INFO

Keywords:
Brain-age estimation
Healthy aging
Machine learning
Regression analysis

ABSTRACT

The difference between age predicted using anatomical brain scans and chronological age, i.e., the brain-age delta, provides a proxy for atypical aging. Various data representations and machine learning (ML) algorithms have been used for brain-age estimation. However, how these choices compare on performance criteria important for real-world applications, such as; (1) within-dataset accuracy, (2) cross-dataset generalization, (3) test-retest reliability, and (4) longitudinal consistency, remains uncharacterized. We evaluated 128 workflows consisting of 16 feature representations derived from gray matter (GM) images and eight ML algorithms with diverse inductive biases. Using four large neuroimaging databases covering the adult lifespan (total $N = 2953$, 18–88 years), we followed a systematic model selection procedure by sequentially applying stringent criteria. The 128 workflows showed a within-dataset mean absolute error (MAE) between 4.73–8.38 years, from which 32 broadly sampled workflows showed a cross-dataset MAE between 5.23–8.98 years. The test-retest reliability and longitudinal consistency of the top 10 workflows were comparable. The choice of feature representation and the ML algorithm both affected the performance. Specifically, voxel-wise feature spaces (smoothed and resampled), with and without principal components analysis, with non-linear and kernel-based ML algorithms performed well. Strikingly, the correlation of brain-age delta with behavioral measures disagreed between within-dataset and cross-dataset predictions. Application of the best-performing workflow on the ADNI sample showed a significantly higher brain-age delta in Alzheimer's and mild cognitive impairment patients compared to healthy controls. However, in the presence of age bias, the delta estimates in the patients varied depending on the sample used for bias correction. Taken together, brain-age shows promise, but further evaluation and improvements are needed for its real-world application.

1. Introduction

Precision and preventive medicine, e.g., early detection of Alzheimer's disease (AD), can benefit from individual-level quantification of atypical aging. Machine learning (ML) approaches, together with large neuroimaging datasets can provide such individualized predictions. Indeed, ML algorithms can capture the multivariate pattern of age-related changes in the brain associated with healthy or typical aging (Franke et al., 2010; Varikuti et al., 2018; Cole 2020; Beheshti et al., 2022; Hahn et al., 2022). Such a model can then be used to predict age, i.e., brain-age, from an unseen subject's image. Being a normative model, a large deviation between the chronological and the predicted age is indicative of atypical aging. A higher positive difference between the brain-age and chronological age, i.e., brain-age delta (which we refer to simply as delta), indicates "older-appearing" brains. As an indicator of future risk of experiencing age-associated health issues, delta quantitatively relates to several age-related risk factors and general physical health, such as weaker grip strength, poorer lung function, history of stroke, greater frequency of alcohol intake, increased mortality risk (Cole et al., 2018; Cole, 2020), and poorer cognitive functions such as fluid intelligence, processing speed, semantic verbal fluency, visual attention, and cognitive flexibility (Cole et al., 2018; Boyle et al., 2021; Richard et al., 2018; Gaser et al., 2013; Cole et al., 2017). Overall, the delta can potentially serve as an omnibus biomarker of brain integrity

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

* Corresponding author at: Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany.
E-mail address: k.patil@fz-juelich.de (K.R. Patil).

https://doi.org/10.1016/j.neuroimage.2023.119947.
Received 20 November 2022; Received in revised form 8 February 2023; Accepted 15 February 2023
Available online 16 February 2023.

and health if its reliability, given different ML workflow designs and other analyses, can be established.

Studies have shown global and local gray matter (GM) volume (GMV) loss (Good et al., 2001; Galluzzi et al., 2008; Giorgio et al., 2010) with aging and accelerated loss in neurodegenerative disorders (Good et al., 2001; Karas et al., 2004; Fjell et al., 2014). This makes GMV a clinically relevant candidate for the investigation of atypical aging via brain-age estimation (Franke et al., 2010; Cole et al., 2015). Brain-age prediction models tend to perform better using GMV than white matter volume (WMV) (Cole et al., 2017; Monté-Rubio et al., 2018), making GMV a promising candidate for further investigation. Furthermore, by reducing the methodological and data-related variance in a model's prediction error, the delta can better reflect a biological signal related to atypical aging.

A brain-age estimation workflow consists of a feature space and an ML algorithm, and several choices exist for each. For instance, voxel-wise data with additional smoothing and/or resampling or parcel-wise averages within a brain atlas can be used as features (Varikuti et al., 2018; Eickhoff et al., 2021). Further dimensionality reduction methods such as principal components analysis (PCA) can improve the observations-to-features ratio and signal-to-noise ratio (Franke et al., 2010; Franke et al., 2013; Gaser et al., 2013). One also needs to choose from a large pool of ML algorithms, such as relevance vector regression (RVR), and Gaussian process regression (GPR), many of which have shown success in brain-age estimation. These choices are known to affect performance (Gutierrez Becker et al., 2018; Baecker et al., 2021; de Lange et al., 2022).

Studies using voxel-based morphometry (VBM)-derived GMV to predict brain-age have claimed prediction errors of ~5–8 years in healthy individuals (Table S1). However, it is difficult to compare these studies as they differ in experimental setup and methodology, such as feature space used, ML algorithms, age range, and evaluation criteria. For a brain-age estimation model to be used in real-world applications, it must perform well on several evaluation criteria; (1) a model should generalize well on new data from the training site as well as on data from novel sites, (2) estimated age must be reliable on repeated measurements, and (3) it should also exhibit longitudinal consistency, i.e., the predicted age should be proportionally higher for later scans after a longer duration, assuming no significant change in lifestyle or health-related interventions between the measurements.

A critical aspect, especially for clinical application, is the commonly reported negative correlation between delta and true age (Beheshti et al., 2019; Smith et al., 2019; de Lange and Cole, 2020). This may result in spurious correlations between the delta and non-imaging measures when chronological age is not accounted for (Franke et al., 2013; Löwe et al., 2016). This age bias complicates or may even mislead downstream individualized decision-making. It can be mitigated using bias correction models; usually, linear regression predicting brain-age or delta using chronological age (Le et al., 2018; Liang et al., 2019; Smith et al., 2019; de Lange et al., 2022). The data source (within or cross-data) and size used to obtain bias correction models has substantial impact on quality of the model. Taken together, there is a gap in understanding the impact of the choices in designing brain-age workflows, and how they affect estimation and utility of individual-level delta.

To fill this gap, we systematically assessed 128 workflows consisting of 16 feature spaces derived from GM images and eight ML algorithms with diverse inductive biases. Using several large neuroimaging databases with a wide age range, we first evaluated these workflows for their within-dataset and cross-dataset performances. Next, we evaluated the test-retest reliability and longitudinal consistency of some top-performing workflows. Then, we assessed the performance of our best-performing workflow in a clinical sample. We investigated the correlations between delta and behavioral/cognitive measures in healthy and clinical cohorts and various factors affecting these correlations. We also compared our best-performing workflow with a publicly available model, brainageR. Several follow-up analyses were performed to investigate the effect of preprocessing (CAT vs. SPM) and tissue type (GM vs. GM+WM+CSF) choices on prediction performance. Finally, given recent evidence that lower accuracy models may capture atypical aging better (Bashyam et al., 2020), we investigated relationship of model performance with delta and delta-behavior correlations.

2. Material and methods

2.1. Datasets

2.1.1. MRI data

We used T1-weighted (T1w) magnetic resonance imaging (MRI) data from healthy subjects covering a wide age range (18–88 years, training data) from several large neuroimaging datasets (Table 1), including the Cambridge center for Ageing and Neuroscience (Cam-CAN, N = 651) (Taylor et al., 2017), Information eXtraction from Images (IXI, N = 562) (https://brain-development.org/ixi-dataset/), the enhanced Nathan Kline Institute-Rockland Sample (eNKI, N = 597) (Nooner et al., 2012), the 1000 brains study (1000BRAINS; N = 1143) (Caspers et al., 2014), Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 2014), the Open Access Series of Imaging Studies (OASIS-3) (LaMontagne et al., 2019), and the MyConnectome dataset (Poldrack et al., 2015). The inclusion criteria were age between 18 and 90 years, gender data available, and no current or past known diagnosis of neurological, psychiatric, or major medical conditions. The IXI dataset was acquired from multiple sites; however, we treat it as a single dataset representing typical data acquired in a noisy clinical setting. From the OASIS-3 dataset, we selected scans from healthy control subjects acquired on 3T scanners. Some other datasets used by brainageR were also used for a fair comparison with our best workflow. The corresponding details are provided in the Supplementary Table S8.

We used the Alzheimer's Disease Neuroimaging Initiative (ADNI; https://adni.loni.usc.edu/) database to evaluate the utility of brain-age in neurodegenerative disorders (Jack et al., 2008; Petersen et al., 2010). We included 3T T1w images from healthy control (HC, N = 209), early and late mild cognitively impaired (EMCI, N = 237; LMCI, N = 128), and Alzheimer's disease (AD, N = 125) subjects. For some of these subjects, data were available for the second timepoint 1–2 years apart (HC, N = 153; EMCI, N = 197; LMCI, N = 104; AD, N = 61) (Table 1d).

2.1.2. Non-imaging data

We used various behavioral/cognitive measures to compute their correlations with delta. Fluid intelligence (FI; N = 631) assessed by the Cattell Culture Fair test and reaction time for the motor learning task (N = 302) from the CamCAN dataset (Taylor et al., 2017). From the eNKI dataset, we used the Color-Word Interference Test (CWIT) inhibition trial completion time (N = 340), the Trail Making Test (TMT) number-letter switching condition completion time (N = 344), Wechsler Abbreviated Scale of Intelligence (WASI-II) matrix reasoning scores (N = 347), and WASI-II similarities scores (N = 347) (Nooner et al., 2012).

Three cognitive tests from ADNI measuring disease severity were used; Mini-Mental State Examination (MMSE), Global Clinical Dementia Rating Scale (CDR), and Functional Assessment Questionnaire (FAQ).

All the datasets except the 1000BRAINS data are available publicly. Ethical approval and informed consent were obtained locally for each study covering both participation and subsequent data sharing. The ethics proposals for the use and retrospective analyses of the datasets were approved by the Ethics Committee of the Medical Faculty at the Heinrich-Heine-University Düsseldorf.

2.2. Data preparation

For the main analysis all T1w images were preprocessed using the Computational Anatomy Toolbox (CAT) version 12.8

Table 1
Sample characteristics of the datasets used in the current study. Datasets used a. for training within-dataset models. b. for training cross-dataset models. c. to evaluate test-retest reliability and longitudinal consistency of brain-age delta and comparison with brainageR (note: for CoRR full sample, the demographics are reported for the last iteration). d. to evaluate performance in clinical samples. Abbreviations: CamCAN: the Cambridge center for ageing and Neuroscience, IXI: Information eXtraction from Images (includes 1.5 and 3T scans), eNKI: the enhanced Nathan Kline Institute-Rockland Sample, CoRR: Consortium for Reliability and Reproducibility, OASIS-3: the Open Access Series of Imaging Studies, ADNI: the Alzheimer's Disease Neuroimaging Initiative, HC: healthy control, EMCI and LMCI: early and late mild cognitively impaired, AD: Alzheimer's disease.

a.

Train dataset	No. of subjects (N)	Males/Females	Age range	Mean ± S.D.	Median
CamCAN	651	321/330	18 - 88	54.27 ± 18.58	54.50
IXI	562	249/313	20 - 86	48.70 ± 16.44	48.85
eNKI	597	188/409	18 - 85	48.25 ± 18.51	50.00
1000BRAINS	1143	660/513	22 - 85	61.85 ± 12.39	63.60

b.

Train dataset	Train N	Test dataset	Test N
IXI + eNKI + 1000BRAINS	2302	CamCAN	651
CamCAN + eNKI + 1000BRAINS	2391	IXI	562
IXI + CamCAN + 1000BRAINS	2356	eNKI	597
IXI + CamCAN + eNKI	1810	1000BRAINS	1143
IXI + CamCAN + eNKI + 1000BRAINS	2953	CoRR, OASIS-3, MyConnectome, ADNI	See below (c & d)

c.

Dataset	Data Filtering	N (sessions)	Males/Females	Age Range	Mean ± S.D.	Median
CoRR	Retest < 3 months	86 (2)	39/47	20.0 - 84.0	48.82 ± 18.28	49.00
	Retest 1 – 2 years	95 (2)	52/43	18.0 - 88.0	34.43 ± 22.51	20.00
	Retest 2 – 3.25 years	26 (2)	18/8	18.0 - 57.0	28.09 ± 11.89	24.50
	Full sample	107	51/56	18.0 – 88.0	49.99 ± 18.87	50.00
OASIS-3	Retest < 3 months	36 (2)	21/15	42.66 - 80.90	63.46 ± 8.80	62.93
	Retest 3- 4 years	127 (2)	52/75	46.04 - 86.21	65.59 ± 8.39	65.90
	Full sample	806	338/468	43.00 - 89.00	69.07 ± 9.06	69.00
MyConnectome	Retest < 3 years	1 (20)	1/0	45.39 - 48.02	45.73 ± 0.58	45.56

d.

Dataset	Disease	N	Males/Females	Age Range	Mean ± S.D.	Median
ADNI (Timepoint-1)	HC	209	99/110	56.3 - 94.7	75.67 ± 6.94	75.50
	EMCI	237	128/109	55.7 - 88.7	70.88 ± 7.12	70.40
	LMCI	128	62/65	55.1 - 91.5	72.02 ± 7.89	72.55
	AD	125	65/60	56.0 - 91.0	74.68 ± 7.99	75.40
ADNI (Timepoint-2)	HC	153	70/83	57.3 - 95.8	75.89 ± 6.63	75.50
	EMCI	197	108/89	56.7 - 90.4	71.81 ± 7.04	71.10
	LMCI	104	51/53	56.1 - 92.5	73.36 ± 7.92	73.95
	AD	61	32/29	57.0 - 93.0	75.79 ± 7.83	76.80

(Gaser et al., 2022). To ensure accurate normalization and segmentation, initial affine registration of T1w images was done with higher than default accuracy (accstr = 0.8). After bias field correction and tissue class segmentation, accurate optimized Geodesic shooting (Ashburner and Friston, 2011) was used for normalization (regstr = 1). We used 1 mm Geodesic Shooting templates and outputted 1 mm isotropic images. The normalized GM segments were then modulated for linear and non-linear transformations.

For comparison with the brainageR model, we used the seven datasets used by brainageR (Table S8) and preprocessed them using CAT 12.8 (Section 2.9). To evaluate the effect of preprocessing and tissue types, we used the SPM12 based preprocessing as implemented by brainageR , which outputs three tissue segmentations (GM, WM, and CSF; see https://github.com/james-cole/brainageR/).

2.3. Workflows

Each workflow consists of a feature representation and an ML algorithm. We evaluated 128 workflows constituting 16 feature representations and eight ML algorithms.

2.3.1. Feature representations

The 16 feature representations were derived from the CAT-preprocessed voxel-wise GM images. Using voxel-wise data can lead to overfitting due to the curse of dimensionality owing to a large number of features compared to the number of samples. Hence, we implemented two dimensionality reduction approaches previously used for brain-age prediction.

In the first strategy, we used voxel-wise GMV after smoothing and resampling (Franke et al., 2010), which may also improve the signal-to-noise ratio. In the second strategy, we used an atlas to summarize data from distinct brain regions (called parcels). This resulted in 16 feature representations.

1. SX_RY: A whole-brain mask was used to select 238,955 voxels. Then, smoothing (S) with an X mm FWHM Gaussian kernel and resampling (R) using linear interpolation to Y mm spatial resolution were applied with $X = \{0, 4, 8\}$ and $Y = \{4, 8\}$, resulting in six feature spaces (S0_R4, S0_R8, S4_R4, S4_R8, S8_R4, S8_R8; SX_R4: 29,852 voxels and SX_R8: 3747 voxels).
2. SX_RY + PCA: Additionally, PCA (Jolliffe, 2002) was applied to each SX_RY feature space while retaining 100% variance, creating an-

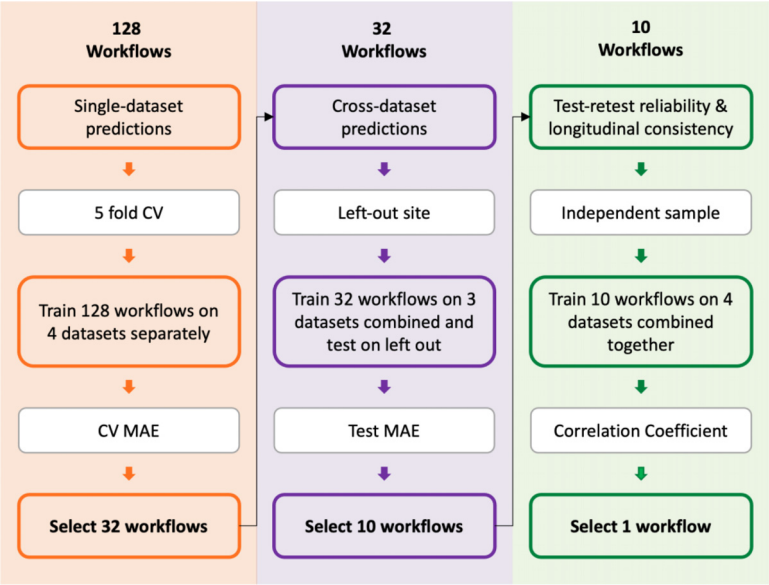


Fig. 1. The framework to select the best-performing workflow for brain-age prediction. A total of 128 workflows were first evaluated for their within-dataset prediction performance using five-fold cross-validation (CV). Next, 32 workflows were selected based on the CV mean absolute error (MAE) and assessed for cross-dataset prediction performance. Within-dataset and cross-dataset evaluations were performed using four datasets (CamCAN, IXI, eNKI and 1000BRAINS). Then, 10 workflows out of 32 were selected based on their test MAE and assessed for test-retest reliability and longitudinal consistency using OASIS-3 and CoRR datasets. The best-performing workflow was selected after considering all the evaluation criteria.

other six representations (S0_R4 + PCA, S0_R8 + PCA, S4_R4 + PCA, S4_R8 + PCA, S8_R4 + PCA, S8_R8 + PCA).

3. Parcel-wise: Four parcel-wise feature spaces were created by combining cortical {100, 400, 800, 1200} parcels (Schaefer et al., 2018) with 36 subcortical (Fan et al., 2016) and 37 cerebellum (Buckner et al., 2011) parcels. We calculated the mean GMV of all the voxels within each parcel (173, 473, 873, and 1273 features).

2.3.2. Machine learning algorithms

We included eight ML algorithms covering diverse inductive biases: ridge regression (RR), least absolute shrinkage and selection operator (LASSO) regression (LR), elastic net regression (ENR), kernel ridge regression (KRR), random forest regression (RFR), GPR, RVR with the linear kernel (RVRlin), and polynomial kernel of degree 1 (RVRpoly). These algorithms have been previously used in the prediction of age and other behavior variables from neuroimaging data (Franke et al., 2010; Gaser et al., 2013; Su et al., 2013; Cole et al., 2015; Varikuti et al., 2018; Jonsson et al., 2019; Liang et al., 2019; Zhao et al., 2019; He et al., 2020; Baecker et al., 2021; Boyle et al., 2021; Lee et al., 2021; Peng et al., 2021; Treder et al., 2021; Vidal-Pineiro et al., 2021; Beheshti et al., 2022; Cole, 2020) (Table S1). Details of these algorithms are provided in the Supplementary Methods.

Recently, deep-learning (DL) models have been applied for brain-age estimation with success (Jiang et al., 2019; Jonsson et al., 2019; Peng et al., 2021). However, in this work, we focus on conventional ML models for the following reasons: (1) ML models have shown competitive performance to DL models (Cole et al., 2017; He et al., 2020; Schulz et al., 2020; Grinsztajn et al., 2022), and (2) the resources required for ML are more readily available and thus still enjoy wider applicability with a lower computational footprint (Thompson et al., 2020; van Wynsberghe, 2021).

2.3.3. Learning setup and software

The ML algorithm's hyperparameters were estimated in a nested fashion using an inner cross-validation (CV) (Varoquaux et al., 2017). Before training, features with low variance were removed (threshold < 1e-5), and the remaining features were Z-scored to have zero mean and unit variance. Any preprocessing steps, including PCA, were applied in a CV-consistent fashion to avoid data leakage, i.e., the parameters were estimated on the training set and applied to both the training and the test set (More et al., 2021).

All the workflows were implemented in Python version 3.9.1 using the Julearn machine-learning library (https://juaml.github.io/julearn/), which in turn uses the scikit-learn library for the learning algorithms KRR, GPR, and RFR (http://scikit-learn.org/) (Pedregosa et al., 2011). LR, RR, and ENR were implemented using the Python wrapper for glmnet (https://pypi.org/project/glmnet/) (Friedman et al., 2010). RVRlin and RVRpoly were implemented using the scikit-rvm package (https://github.com/JamesRitchie/scikit-rvm/). The codes used for preprocessing, feature extraction, model training and prediction are available at https://github.com/juaml/brainage_estimation.

2.4. Analysis setup

Given data acquisition and site-related biases, it is important to identify a workflow that shows high accuracy in different evaluation scenarios. For instance, a workflow that works well on a dataset might not work well on another dataset. To accommodate such real-world scenarios, we followed a systematic procedure where the workflows were subjected to increasingly stringent evaluations (Fig. 1). In brief, we first evaluated the within-dataset CV performance of the 128 workflows. Next, 32 workflows characterizing the overall pattern of performance were selected for cross-dataset evaluation. This selection was performed by uniformly sampling over the within-dataset CV performance. This allows for the possibility that workflows with low within-dataset performance might perform well in cross-dataset evaluation. Finally, the top 10 workflows out of the 32 were evaluated for their test-retest reliability and longitudinal consistency. After considering all the evaluation criteria, the best-performing workflow was chosen and used for application on ADNI data and comparison with brainageR. Specific analysis steps are described below.

2.4.1. Within-dataset and cross-dataset evaluations

We evaluated the 128 workflows (see Section 2.3) separately on four datasets, CamCAN, IXI, eNKI, and 1000BRAINS. This scenario assumes that enough within-dataset training data are available and is widely used in brain-age estimation work (Ashburner, 2007; Su et al., 2013; Gutierrez Becker et al., 2018). To estimate a single out-of-sample brain-age for each subject, we used a 5-fold CV. For each hold-out (test) fold, the remaining 80% of the data were used for training and to obtain a generalization estimate using 5 times repeated 5-fold (5 × 5-fold) nested CV. All CV analysis was stratified by age to preserve the age distribution. It is important to obtain a single prediction per subject (as opposed to

multiple predictions per subject if the outer CV were repeated) for further meaningful analyses, such as correlation with non-imaging measures. Consequently, we computed two measures, test performance, and CV performance. The test performance was obtained by averaging over the outer 5 folds. The CV performance was obtained by first averaging over the inner 5 × 5-fold CV and then over the outer 5-fold CV. Finally, the CV and test performance were averaged over the four datasets. The performance was evaluated using mean absolute error (MAE), Pearson's correlation between predicted and true (chronological) age, and the coefficient of determination R^2.

We followed a systematic procedure to select a subset of workflows while maintaining diversity in terms of CV performance. Specifically, the workflows were arranged in the increasing order of their average CV MAE and divided into 16 groups. Next, the top two workflows (with the lowest CV MAE) from each group were selected.

We tested these 32 selected workflows on cross-dataset to obtain sample-unbiased performance. This emulates the real-world scenario where data from the application site are not available, and the training and test data come from different sources with confounding effects, such as scanner hardware or operator inconsistencies (Jovicich et al., 2006; Chen et al., 2014). Three out of four datasets (CamCAN, IXI, eNKI and 1000BRAINS) were pooled to form the training data, and the hold-out dataset was used as the test data. A 5 × 5-fold CV was performed on the training data to estimate the generalization performance with an internal CV for hyperparameter tuning. The CV performance was averaged over 5 × 5-fold CV and then over the four hold-out datasets. The test performance was averaged over the four datasets. The performance was again evaluated using MAE, Pearson's correlation between predicted and true age, and the coefficient of determination R^2.

The 32 workflows were arranged in increasing order of their average test MAE, i.e., their average performance on the hold-out datasets, from which the top 10 workflows were selected.

2.4.2. Test-retest reliability and longitudinal consistency

We then trained models using the 10 selected workflows with the four datasets combined as training data (IXI + eNKI + CamCAN + 1000BRAINS, $N = 2953$; Supplementary Fig. S1). The test-retest reliability and longitudinal consistency of the delta were evaluated for the 10 models using the OASIS-3 and CoRR datasets.

To evaluate test-retest reliability, we used: two scans from the same subjects acquired within a delay of (1) less than three months (CoRR: $N = 86$, age range = 20–84 years, OASIS-3: $N = 36$, age range = 43–81), and (2) between 1 and 2 years (CoRR: $N = 95$, age range = 18–88). The concordance correlation coefficient (CCC) (Lin, 1989) between the delta (predicted age minus age at the scan time) from the two scans was calculated.

To evaluate longitudinal consistency, two scans from the same subjects acquired with a retest duration (1) between 2 and 3.25 years (CoRR: $N = 26$, age range = 18–57), and (2) between 3 and 4 years (OASIS-3: $N = 127$, age range = 46–86) were used. We computed Pearson's correlation between the difference in the predicted age and the difference in chronological age from the two scans. A higher positive correlation here would indicate higher longitudinal consistency.

By considering the results from the within- and cross-dataset analysis, test-retest reliability, and longitudinal consistency, we chose one best-performing workflow for further analysis.

2.5. Bias correction

Many studies have reported age-dependency of the delta with over-prediction in young subjects and under-prediction in older subjects (Le et al., 2018; Liang et al., 2019), which renders the usage of delta as an individualized biomarker problematic. A common practice is to apply a statistical bias correction to remove the effect of age from either the predicted age or the delta (Le et al., 2018; Liang et al., 2019; Smith et al., 2019; Cole, 2020; de Lange and Cole, 2020). Note that when calculating correlations of delta with non-imaging measures, bias correction is expected to be similar to partial correlation analysis when age is used as a covariate.

Several alternatives are available for bias correction (de Lange et al., 2019; Cole, 2020; de Lange and Cole, 2020; Smith et al., 2019(Beheshti et al., 2019)). We chose the method used by Cole and colleagues (Cole, 2020) as it does not use the chronological age of the test data, and thus avoids information leakage which can bias comparison between workflows by making low-performing workflows appear good (de Lange et al., 2022). Furthermore, this method is relevant for possible future applications like forensic investigations where test age is not available. A linear regression model was fitted with the out-of-sample (from the CV) predicted age as the dependent variable and chronological age as the independent variable using the training data. The predicted age in the test set was corrected by subtracting the resulting intercept and dividing by the slope.

2.6. Correlation with cognitive measures

To understand the effect of bias correction and the impact of covariates on delta-behavior correlations, we performed correlations of behavior/cognitive measures from CamCAN and eNKI datasets (see Section 2.1.2) with (1) uncorrected delta, (2) uncorrected delta with age as a covariate, (3) corrected delta, and (4) corrected delta with age as a covariate. If the bias correction eliminates the antagonistic relation between delta and age, we expect (2), (3), and (4) to give similar correlations. Furthermore, to assess the impact of data used for learning bias correction models, we performed these analyses using delta obtained from within-dataset and cross-dataset predictions.

2.7. Brain-age in clinical samples

Next, we used the ADNI dataset (Jack et al., 2008; Petersen et al., 2010) to validate our best-performing workflow on clinical samples. We estimated and compared the delta between HC, EMCI, LMCI, and AD subjects (Table 1d).

Our best-performing workflow trained on the four datasets was used to obtain the predictions, followed by application of bias correction model (see Section 2.5). We compared two bias correction models, one derived using the CV predictions from the four training datasets and another using HC samples in ADNI data (Franke and Gaser, 2012). The group-wise corrected delta was compared using analysis of variance (ANOVA) followed by Bonferroni correction to counteract multiple comparisons. Emulating the scenario that application sites might have different numbers of HC samples, we learned bias correction models using HC sub-samples (0.1 to 0.9 fraction in steps of 0.1) drawn without replacement and applied them on the full HC and AD samples. This process was repeated 100 times to estimate the variance of mean corrected delta in HC and AD subjects.

Finally, we investigated associations between the corrected delta and three clinical test scores, MMSE, CDR, and FAQ. The correlations were computed using the whole sample and different diagnostic groups separately using Pearson's correlation with age as a covariate for both sessions separately.

2.8. Relationship of MAE with delta and delta-behavior correlations

Here, we sought to select a workflow that provides accurate and reliable predictions. We reason that a workflow that accurately predicts the age of healthy individuals captures the typical brain aging process, and thus, a large delta in new data can be considered indicative of atypical aging. However, recent evidence shows that an overfitted brain-age model (high training accuracy) is not the most sensitive in identifying pathologies (Bashyam et al., 2020). This study showed that a relatively moderately fit model yielded brain-age deltas with more sig-

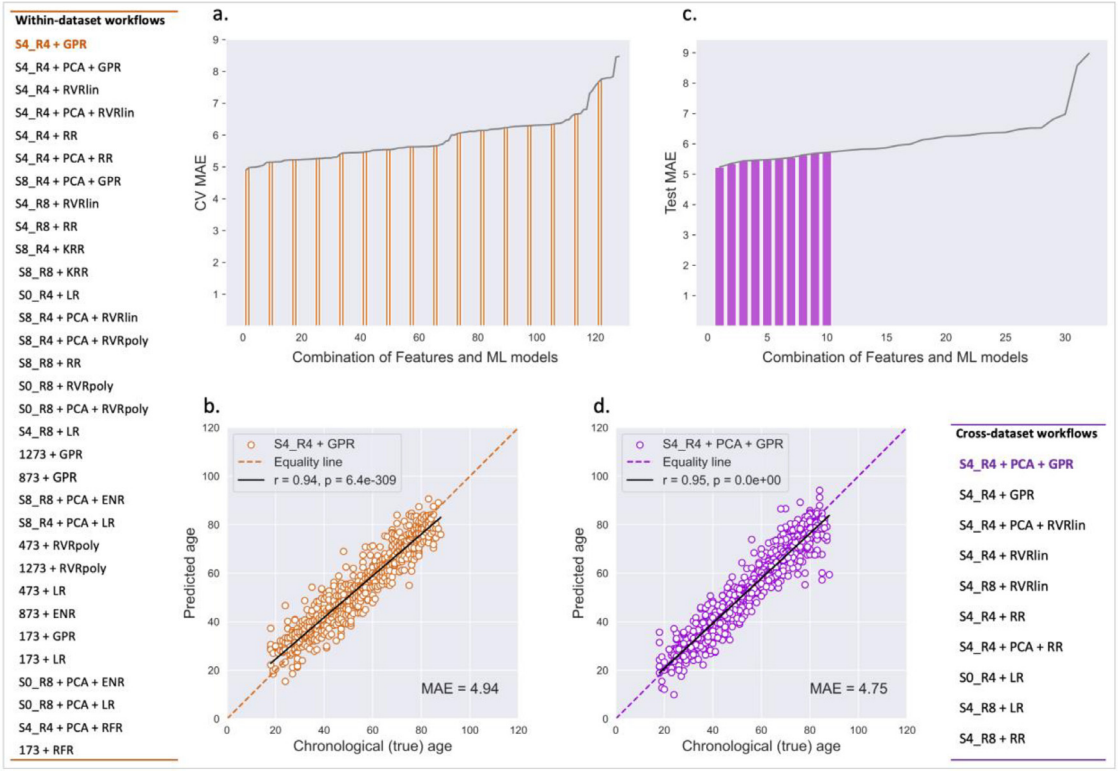


Fig. 2. Within-dataset and cross-dataset results. a. The line plot showing CV MAE (averaged across four datasets) for 128 workflows arranged in increasing order (names of all workflows are given in Table S2). The orange bars represent the MAEs of 32 selected workflows with their names in the table on left. b. The scatter plot between the chronological age and within-dataset predicted age for the CamCAN data using S4_R4 + GPR workflow (MAE = 4.94 years and $r = 0.94$, $p = 6.4e\text{-}309$). c. The line plot showing test MAE (averaged across four runs) for the 32 workflows arranged in increasing order (names of all workflows are given in Table S3). The purple bars represent the MAEs of 10 selected workflows with their names in the table on the bottom right. d. The scatter plot between the chronological age and cross-dataset predicted age for the CamCAN data using S4_R4 + PCA + GPR workflow (MAE = 4.75 years and $r = 0.95$, $p = 0.0e+00$).

nificant group differences and the larger effect sizes between control and disease groups in various brain pathologies.

To investigate this possibility, we trained the 32 workflows selected from the cross-dataset analysis with four datasets pooled together for training and applied to timepoint 2 ADNI data. To understand how the model performance varies with its utility, we compared the models' MAEs with the corrected mean delta in AD sample and examined whether it was related to the delta-behavior correlations. We then performed a similar analysis in two HC samples (CamCAN and eNKI) using corresponding within-dataset hold-out predictions.

2.9. Comparison with brainageR and effect of preprocessing and tissue types

We compared the performance of our best-performing workflow with an already available brain-age estimation model, brainageR. The brainageR model was trained on 3377 healthy individuals (age range = 18–92 years, mean ± SD age = 40.6 ± 21.4 years) from seven publicly available datasets using the GPR algorithm. It uses SPM12 to segment and normalize T1w images, from which GM, WM, and CSF vectors were extracted (using 0.3 probability masked brainageR-specific templates). PCA was used to reduce data dimensionality, and 435 components explaining 80% of the variance were retained. Note that brainageR uses three tissue types, while our focus is on GM.

To avoid bias due to different training data, for this comparison we used data from the same subjects used by brainageR (2 subjects could not be processed; Table S8). Next, using this training data, we trained our best-performing workflow using GMV extracted from CAT 12.8 and compared the performance with already trained brainageR model on three datasets, (1) CoRR (N = 107, sub-sampled to keep uniform distribution in age-range = 18–88 years, repeated 100 times; see Supplementary Methods for more details), (2) the OASIS-3 (N = 806; first scan per subject, age-range = 43–89 years), and (3) the MyConnectome study (one subject scanned 20 times in a period of 3 years; age range = 45–48 years). Additionally, we used sub-samples from OASIS-3 with test-retest durations of (1) less than 3 months (N = 36, 43–81 years) and (2) between 3 and 4 years (N = 127, 46–86 years) to evaluate test-retest reliability and longitudinal consistency, respectively (see Section 2.4.2).

Next, we compared how the preprocessing and tissue types affect model performance. Following our focus on GMV, we compared; (1) CAT-preprocessed GMV, (2) SPM-preprocessed GMV, and (3) SPM-preprocessed GM, WM, and CSF images following brainageR. The latter investigates whether WM and CSF features provide complementary information leading to better predictions. For this, we performed within-dataset evaluation on IXI and CamCAN datasets (see Section 2.4.1).

3. Results

3.1. Within-dataset and cross-dataset predictions

For within-dataset analysis, the CV performance (average over 125 estimates–inner 5 × 5-fold CV, repeated 5 times, see Section 2.4.1) and test performance based on single prediction per subject from the outer CV, were calculated. These were then averaged separately over four datasets.

The average CV MAE (4.90–8.48 years) and the average test MAE (4.73–8.38 years) (Fig. 2a, Table S2) were similar, indicating that the nested CV generalization estimates are indeed indicative of their test performance. The correlation between the true and predicted age on the test data ranged from 0.81 to 0.93, while the age bias (correlation

Table 2
The performance metric for the best workflow on different datasets. A. Within-dataset prediction (using S4_R4 + GPR) b. Cross-dataset prediction (using S4_R4 + PCA + GPR). Abbreviations: MAE: mean absolute error between true and predicted age, MSE: mean squared error between true and predicted age, R^2: the proportion of variance of predicted age explained by the independent variables in the model, Corr (true, pred): Pearson's correlation between true and predicted age, Age bias: Pearson's correlation between true age and brain-age delta.

Datasets	N	a. Within-dataset results					b. Cross-dataset results				
		MAE	MSE	R^2	Corr (true, pred)	Age bias	MAE	MSE	R^2	Corr (true, pred)	Age bias
CamCAN	651	4.94	39.54	0.89	$r = 0.94$, $p = 6.4e\text{-}309$	$r = -0.42$, $p = 6.8e\text{-}29$	4.75	38.35	0.89	$r = 0.95$, $p = 0.0e{+}00$	$r = -0.23$, $p = 3.1e\text{-}09$
IXI	562	4.76	35.20	0.87	$r = 0.93$, $p = 2.9e\text{-}252$	$r = -0.48$, $p = 3.5e\text{-}33$	6.08	57.35	0.79	$r = 0.94$, $p = 1.2e\text{-}267$	$r = -0.18$, $p = 2.2e\text{-}05$
eNKI	597	5.20	44.85	0.87	$r = 0.93$, $p = 8.1e\text{-}267$	$r = -0.47$, $p = 1.4e\text{-}33$	4.97	39.65	0.88	$r = 0.94$, $p = 9.7e\text{-}288$	$r = -0.49$, $p = 3.6e\text{-}38$
1000- BRAINS	1143	4.04	26.65	0.83	$r = 0.91$, $p = 0.0e{+}00$	$r = -0.50$, $p = 2.0e\text{-}73$	5.13	41.03	0.73	$r = 0.90$, $p = 0.0e{+}00$	$r = -0.15$, $p = 2.0e\text{-}07$

Table 3
Concordance correlation coefficient (CCC) between brain-age delta from two sessions at different test-retest durations and their respective mean absolute error (MAE) between true and predicted age for CoRR and OASIS-3 datasets for the top 10 workflows.

	CoRR dataset						OASIS-3 dataset		
Retest duration Age range (years)	< 3 months (N = 86; 20.0 - 84.0)			1 – 2 years (N = 95; 18.0 - 88.0)			< 3 months (N = 36; 42.66 - 80.90)		
Workflows	MAE (ses-1)	MAE (ses-2)	CCC	MAE (ses-1)	MAE (ses-2)	CCC	MAE (ses-1)	MAE (ses-2)	CCC
S4_R4 + PCA + GPR	*4.808*	*5.008*	0.97	*4.374*	*4.204*	0.95	*4.2*	*3.801*	0.80
S4_R4 + GPR	4.928	5.112	0.97	4.738	4.49	0.96	4.24	3.935	0.82
S4_R4 + PCA + RVRlin	5.811	5.757	0.97	5.156	5.072	0.96	5.288	5.223	0.83
S4_R4 + RVRlin	5.815	5.76	0.97	5.141	5.065	0.96	5.234	5.177	0.83
S4_R8 + RVRlin	6.375	6.265	0.95	5.444	5.33	0.96	5.109	5.2	0.77
S4_R4 + RR	5.64	5.653	*0.98*	5.174	5.277	*0.97*	4.918	4.71	0.85
S4_R4 + PCA + RR	5.742	5.732	*0.98*	5.288	5.404	*0.97*	4.988	4.744	0.85
S0_R4 + LR	6.281	6.359	0.96	6.251	6.293	0.94	4.949	5.161	*0.86*
S4_R8 + LR	6.763	6.676	0.97	6.497	6.434	*0.97*	5.811	5.896	0.79
S4_R8 + RR	6.232	6.185	0.97	5.975	6.016	*0.97*	5.332	5.328	0.81

between true age and delta) ranged from −0.22 to −0.83 (Table S2). Overall, all workflows showed a high similarity in their predictions (correlations 0.83–0.99 averaged across the four datasets; Fig. S2). The top 20 workflows showed comparable CV and test MAE with a difference of less than 0.4 years.

Well-performing workflows primarily consisted of voxel-wise smoothed and resampled feature spaces with and without PCA, with S4_R4 (smoothed with a 4 mm FWHM kernel and resampled to 4 mm spatial resolution) generally performing better. Some workflows with PCA performed similarly to their respective non-PCA version but not all (see Supplementary Table S2). GPR, KRR, RR, and both RVR algorithms generally ranked high. Most algorithms performed worse with parcel-wise features, while RFR generally exhibited the worst performance.

The workflow S4_R4 + GPR performed the best (see Table 2a for its performance on each of the four datasets). This workflow showed the lowest average CV MAE with a high R^2 and a high correlation between true and predicted age (Fig. 2b) but a relatively high age bias (Fig. S3). The second-best workflow, S4_R4 + PCA + GPR, performed similarly to the best workflow. Other workflows with the S4_R4 feature space, with or without PCA, together with the KRR, RVRpoly, and RVRlin algorithms, performed comparably. From the 128 workflows, we selected 32 workflows while preserving diversity in terms of CV MAE.

The 32 workflows selected for cross-dataset analysis showed the average CV (5 × 5-fold on training data) MAE (4.28–7.39 years) lower than the test (hold-out dataset) MAE (5.23–8.98 years) (Fig. 2c). The test-set correlation between true and predicted age ranged from 0.82 to 0.93, while the age bias ranged from −0.27 to −0.75 (Table S3). All workflows showed a high similarity in their predictions (correlations 0.83–0.99 averaged across the four runs). Due to this high similarity, the averaged predictions, i.e., ensemble, from 32 workflows were not better than the top-performing workflow (Fig. S2). The workflows that performed well within-dataset also performed well in cross-dataset predictions (Fig. S6). These results indicate that the corresponding models could generalize well to data from a new unseen site.

We selected 10 workflows with the lowest test MAE for further analysis. These workflows consisted of only voxel-wise feature spaces (S4_R4, S4_R8, and S0_R4) with and without PCA. The ML algorithms included GPR, RVRlin, RR, and LR. The best-performing workflow was the S4_R4 + PCA + GPR with the lowest average test MAE, a high R^2, a high correlation between true and predicted age (Fig. 2d), and moderate age bias (Fig. S3), see Table 2b for its performance on all four datasets), followed by the S4_R4 + GPR workflow.

3.3. Test-retest reliability and longitudinal consistency

The test-retest reliability and longitudinal consistency of the top 10 workflows selected from the cross-dataset evaluation were evaluated using the CoRR and OASIS-3 datasets.

For the short retest duration of less than three months, all 10 workflows showed high test-retest reliability (CoRR: CCC = 0.95–0.98, age range = 20–84 years; OASIS-3: CCC = 0.77–0.86, age range = 43–81 years). For the longer retest duration of 1–2 years in the CoRR dataset, CCC ranged between 0.94–0.97 (age range = 18–88 years) (Table 3). These results show that the age was reliably estimated by the selected workflows.

Next, we evaluated the longitudinal consistency as the correlation between the difference in the predicted age and the difference in the chronological age (Fig. 3, Table S4). Six workflows out of 10 showed a significant positive linear relationship at the retest duration of 2–3.25 years (r between 0.451–0.437, $p < 0.05$) in the CoRR dataset. These workflows included the S4_R4 feature space with and without PCA with the GPR, RVRlin, and RR algorithms. In contrast, none of the workflows showed a linear relationship in the OASIS-3 dataset (retest duration 3–4 years).

Although the workflows showed similar test-retest reliability and longitudinal consistency, the workflow S4_R4 + PCA + GPR showed the lowest MAE on these sub-samples (Tables 3, S4). Therefore, considering all the analysis scenarios, within-dataset, cross-dataset, test-retest reliability, and longitudinal consistency, although other workflows were also competitive, we deemed the S4_R4 + PCA + GPR workflow as well-performing and chose it for further analysis.

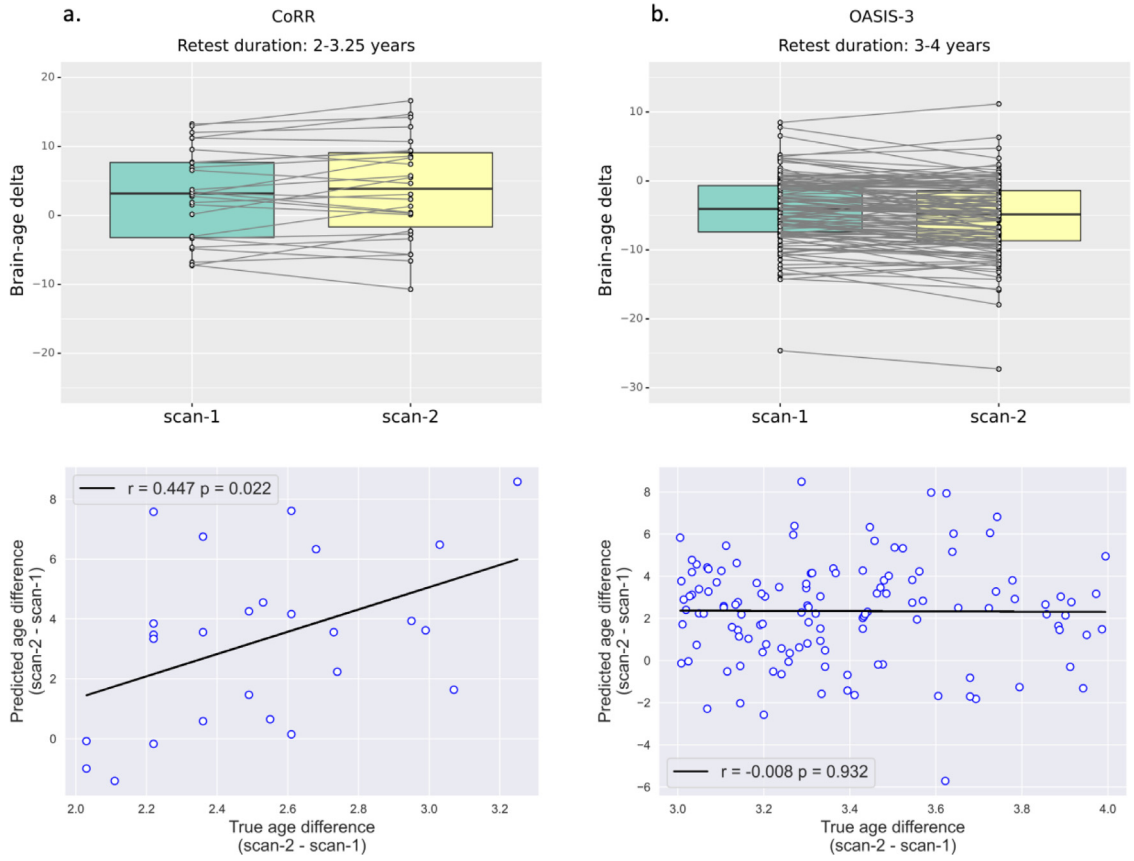


Fig. 3. Longitudinal consistency. (top) The brain-age delta from two scans of the same subjects and (bottom) the scatter plot between the difference in chronological age and the difference in predicted age between two scans acquired within a retest duration of a. 2–3.25 years (CoRR dataset) b. 3–4 years (OASIS-3 dataset).

3.4. Bias correction and correlation with behavioral/cognitive measures

In the CamCAN data, FI was negatively correlated with age ($r = -0.661$, $p = 1.92e\text{-}80$), while motor learning reaction time was positively correlated with age ($r = 0.544$, $p = 1.11e\text{-}24$). In the eNKI data, CWIT inhibition trial completion time ($r = 0.361$, $p = 6.50e\text{-}12$) and TMT number-letter switching trial completion time ($r = 0.279$, $p = 1.45e\text{-}07$) were positively correlated with age. On the other hand, WASI matrix reasoning scores were negatively correlated ($r = -0.240$, $p = 6.03e\text{-}06$), and WASI similarities scores were not correlated ($r = 0.052$, $p = 0.332$) with age (Table 4).

As several ways have been proposed to obtain the correlation between delta and behavior, e.g., using bias-corrected delta or using age as a covariate, we evaluated several alternatives (see Section 2.6).

3.4.1. Within-dataset predictions

Within-dataset hold-out predictions, i.e., single prediction per subject, were derived using the chosen workflow (S4_R4 + PCA + GPR). The bias correction model was estimated using the CV predictions on the same dataset. In both datasets, there was no residual age bias after bias correction: CamCAN, $r = -0.17$, $p = 1.13e\text{-}05$ and $r = 0.00$, $p = 0.999$; and eNKI, $r = -0.20$ $p = 4.53e\text{-}07$ and $r = 0.001$, $p = 0.986$, before and after correction, respectively (Fig. S3).

We first calculated the correlation between the uncorrected delta and behavioral measures using age as a covariate (Table 4a). In the CamCAN data, a higher delta was associated with lower FI ($r = -0.154$, $p = 0.0001$) and higher motor learning reaction time ($r = 0.181$, $p = 0.002$). In the eNKI data, a higher delta was associated with lower response inhibition and selective attention, as indicated by a higher CWIT inhibition trial completion time ($r = 0.109$, $p = 0.045$). There were no correlations between delta and intelligence scores (WASI matrix reasoning and similarities). The results with age, age^2, and gender as covariates showed a similar trend (Table S5a).

Next, we repeated this analysis with the corrected delta (Table 4a) and expected results similar to using uncorrected delta with age as a covariate. We indeed found similar correlations with FI ($r = -0.157$ $p = 7.24e\text{-}05$) and motor learning reaction time ($r = 0.186$ $p = 0.001$) in the CamCAN data, but no significant correlation with CWIT inhibition trial completion time ($r = 0.094$, $p = 0.084$) in the eNKI data. The correlations using corrected delta with covariate were highly similar to uncorrected delta with covariate (Table 4a).

3.4.2. Cross-dataset predictions

Cross-dataset predictions were derived for the CamCAN and eNKI datasets using the S4_R4 + PCA + GPR workflow trained on the IXI + eNKI + 1000BRAINS ($N = 2302$) and IXI + CamCAN + 1000BRAINS ($N = 2356$) datasets, respectively.

In the CamCAN data, the bias correction model was successful with age bias before and after correction $r = -0.23$, $p = 3.06e\text{-}09$ and $r = -0.04$, $p = 0.263$, respectively. However, the correction was not successful in the eNKI data; the age bias was $r = -0.49$, $p = 3.62e\text{-}38$ and $= -0.35$, $p = 8.39e\text{-}19$ before and after correction, respectively (Fig. S3). This result indicates that the bias correction might not always work well when applied to cross-dataset.

Using age as a covariate on the uncorrected delta, we did not find a significant delta-behavior correlation in the CamCAN data. In the eNKI data, a higher delta was associated with lower response inhibition and selective attention, as indicated by a higher CWIT inhibition trial completion time ($r = 0.208$, $p = 0.0001$) and lower cognitive flexibility indicated by a higher TMT completion time ($r = 0.147$, $p = 0.006$) (Table 4b). There were no correlations between delta and intelligence

Table 4
Correlation of brain-age delta with various behavioral measures with and without bias correction. a. From within-dataset predictions. b. From cross-dataset predictions. Age was used as a covariate. Abbreviations: CWIT: Color-Word Interference Test, TMT: Trail Making Test, WASI-II: Wechsler Abbreviated Scale of Intelligence.

a. From within-dataset predictions

Dataset	Behavioral measure	N	Correlation with age	No bias correction		After bias correction	
				(a) No covariate	(b) With covariate	(c) No covariate	(d) With covariate
CamCAN	Fluid Intelligence (Cattel test)	631	$r = -0.661, p = 1.9e\text{-}80$	$r = -0.043, p = 0.282$	**$r = -0.154, p = 0.0001$**	**$r = -0.157, p = 7.2e\text{-}05$**	**$r = -0.154, p = 0.0001$**
	Motor Learning (Reaction time)	302	$r = 0.544, p = 1.1e\text{-}24$	$r = 0.089, p = 0.122$	**$r = 0.181, p = 0.002$**	**$r = 0.186, p = 0.001$**	**$r = 0.180, p = 0.002$**
eNKI	CWIT (Inhibition trial completion time)	340	$r = 0.361, p = 6.5e\text{-}12$	$r = 0.022, p = 0.683$	**$r = 0.109, p = 0.045$**	$r = 0.094, p = 0.084$	**$r = 0.110, p = 0.043$**
	TMT (Number-letter switching trial completion time)	344	$r = 0.279, p = 1.5e\text{-}07$	$r = -0.031, p = 0.564$	$r = 0.033, p = 0.542$	$r = 0.022, p = 0.690$	$r = 0.033, p = 0.545$
	WASI-II matrix reasoning	347	$r = -0.240, p = 6.0e\text{-}06$	$r = 0.026, p = 0.627$	$r = -0.030, p = 0.581$	$r = -0.019, p = 0.728$	$r = -0.029, p = 0.590$
	WASI-II similarities	347	$r = 0.052, p = 0.332$	$r = -0.033, p = 0.536$	$r = -0.022, p = 0.685$	$r = -0.023, p = 0.667$	$r = -0.021, p = 0.698$

b. From cross-dataset predictions

Dataset	Behavioral measure	N	Correlation with age	No bias correction		After bias correction	
				(a) No covariate	(b) With covariate	(c) No covariate	(d) With covariate
CamCAN	Fluid Intelligence (Cattel test)	631	$r = -0.661, p = 1.9e\text{-}80$	$r = 0.071, p = 0.074$	$r = -0.073, p = 0.066$	$r = -0.053, p = 0.180$	$r = -0.073, p = 0.066$
	Motor Learning (Reaction time)	302	$r = 0.544, p = 1.1e\text{-}24$	$r = -0.023, p = 0.689$	$r = 0.092, p = 0.110$	$r = 0.083, p = 0.151$	$r = 0.092, p = 0.110$
eNKI	CWIT (Inhibition trial completion time)	340	$r = 0.361, p = 6.5e\text{-}12$	$r = 0.005, p = 0.931$	**$r = 0.208, p = 0.0001$**	$r = 0.065, p = 0.230$	**$r = 0.208, p = 0.0001$**
	TMT (Number-letter switching trial completion time)	344	$r = 0.279, p = 1.5e\text{-}07$	$r = -0.007, p = 0.898$	**$r = 0.147, p = 0.006$**	$r = 0.039, p = 0.469$	**$r = 0.147, p = 0.006$**
	WASI-II matrix reasoning	347	$r = -0.240, p = 6.0e\text{-}06$	$r = 0.077, p = 0.150$	$r = -0.045, p = 0.400$	$r = 0.043, p = 0.425$	$r = -0.045, p = 0.400$
	WASI-II similarities	347	$r = 0.052, p = 0.332$	$r = -0.098, p = 0.068$	$r = -0.083, p = 0.122$	$r = -0.096, p = 0.073$	$r = -0.083, p = 0.122$

scores (WASI matrix reasoning and similarities). The results with age, age^2, and gender as covariates showed a similar trend (Table S5b).

Since there was a residual correlation between corrected delta and age, the correlations with behavior without age as a covariate can be unreliable. We, therefore, do not discuss correlations of the corrected delta without age as a covariate, but they are reported in Table 4 for completeness. Additionally, as expected, the correlations using corrected delta with age as a covariate were similar to uncorrected delta with covariate (Table 4b).

3.5. Predictions in the ADNI sample

At timepoint 1, the mean uncorrected delta was −5.97 years in HC, −4.39 in EMCI, −3.57 in LMCI, and −2.13 in AD (Fig. 4a). In other words, the model underestimated age. The slope and intercept derived from the bias correction model using the training data (CV predictions) could not entirely correct for the under-estimation and age bias (Fig. 4b). Bias correction using the whole ADNI HC sample removed the bias (average delta, HC = 0, EMCI = 0.85, LMCI = 2.09, AD = 4.47 years) (Fig. 4c). ANOVA revealed that the corrected delta differed significantly across the groups ($F = 12.94$, $p = 3.10e\text{-}08$), and post-hoc t-tests revealed significant differences between AD and HC ($p = 1.16e\text{-}08$), EMCI ($p = 1.87e\text{-}05$), LMCI ($p = 0.043$), and HC and LMCI ($p = 0.022$) after Bonferroni correction. At timepoint 2, the pattern was similar to timepoint 1 but with higher corrected delta values (EMCI = 1.15 years, LMCI = 2.88, AD = 6.59 years) (Fig. 4e-f, Table 5). These results demonstrate that our model could capture the range of normal structural variation related to age in healthy subjects and deviance in both MCI and AD patients.

The correlations between HC sample-corrected delta and various clinical test scores were calculated with age as a covariate (Table 6). At timepoint 1, the delta was negatively correlated with MMSE ($r = -0.255$, $p = 0.016$) and positively correlated with FAQ ($r = 0.275$, $p = 0.005$) in the entire sample. No correlations were found in individual diagnostic groups or could not be calculated due to insufficient score data. At timepoint 2, the delta was negatively correlated with MMSE ($r = -0.303$, $p = 2.40e\text{-}12$) and positively correlated with CDR ($r = 0.270$, $p = 7.35e\text{-}10$) and FAQ ($r = 0.331$, $p = 2.31e\text{-}14$) in the whole sample. In the AD group, the delta was positively correlated with FAQ ($r = 0.298$, $p = 0.021$) but not with MMSE or CDR. In the LMCI group, the delta was positively correlated with FAQ ($r = 0.309$, $p = 0.002$), negatively correlated with MMSE ($r = -0.227$, $p = 0.022$), and not correlated with CDR. In the EMCI group, the delta positively correlated with CDR ($r = 0.153$, $p = 0.034$) but not MMSE and FAQ scores. No correlations were found in the HC group. The correlations with age, age^2, and gender as covariates were similar (Table S6).

We also found that the size of HC sample used for bias correction considerably impacts the mean corrected delta in AD subjects (Fig. S7). Specifically, with fewer HC subjects, the variance of the corrected delta in AD was much higher in both sessions, e.g., at the timepoint 1 when using 21 HC samples, the mean AD delta ranged between ~1–12 years and converged to 4.47 years as the sub-samples approached the complete sample.

3.6. Relationship of MAE with delta and delta-behavior correlations

Using 32 workflows selected from the cross-dataset evaluation, we analyzed whether model performance (MAE) was associated with their brain-behavior correlations. The corrected mean delta in AD ranged from 5.43 to 10.01 years, with some relatively poor performing models yielding a higher delta in AD (Table S7). Lower accuracy (higher MAE) was associated with stronger delta-MMSE correlation (Fig. 5c). In contrast, lower MAE was associated with a stronger brain-behavior correlations in the two healthy samples, delta-motor learning reaction time in CamCAN, and delta-CWIT inhibition trial completion time in eNKI datasets (Fig. 5a & b).

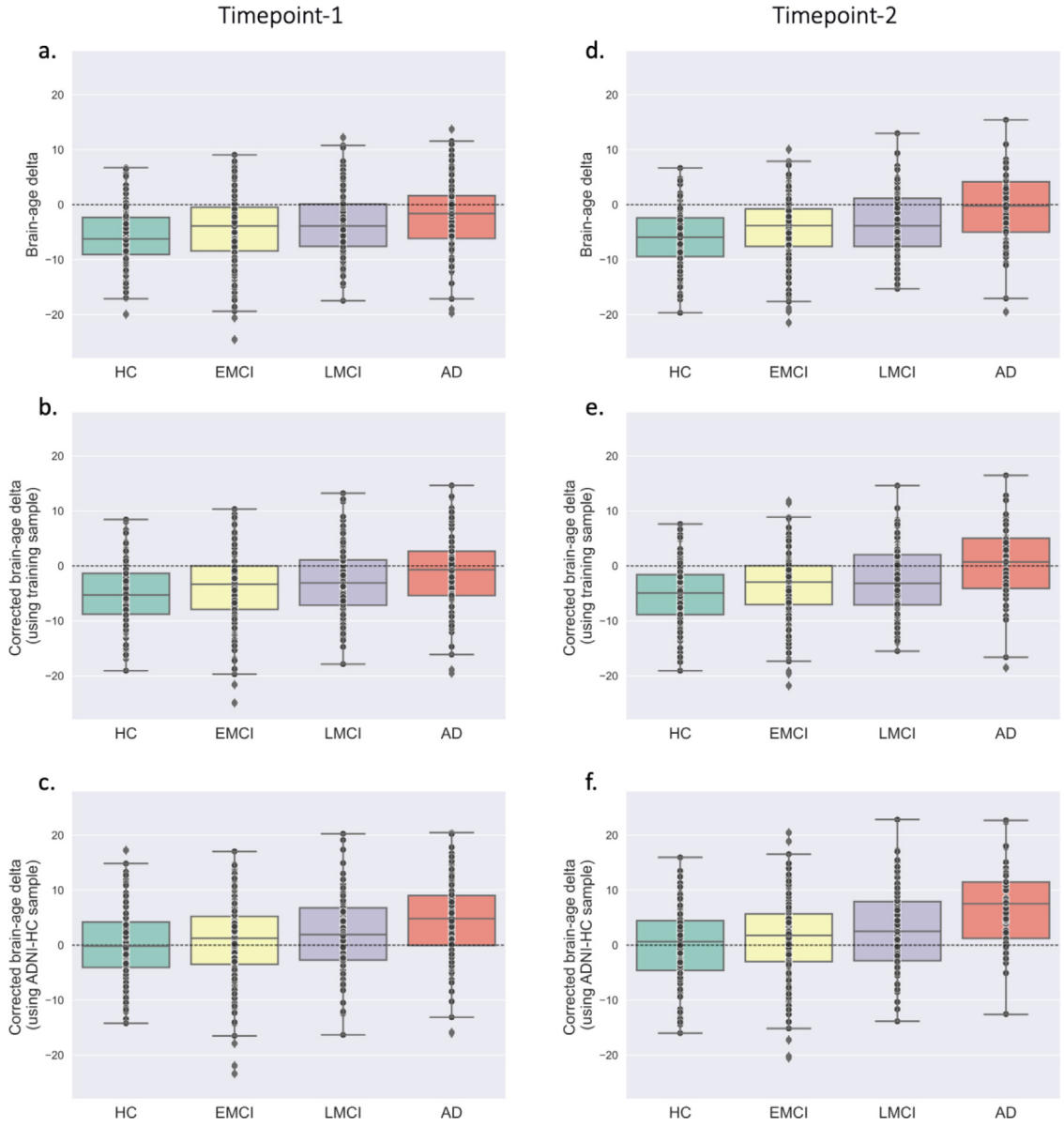


Fig. 4. Brain-age delta in the clinical population. The box plot compares the delta between healthy control (HC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and Alzheimer's disease (AD) from the ADNI sample at (left) timepoint-1 and (right) timepoint-2. Box plot with a & d. uncorrected delta. b & e. corrected delta using the CV predictions from the training set. c & f. corrected delta using the predictions from HC-ADNI subjects.

Table 5
Prediction performance on the ADNI data from two timepoints using the best-performing (S4_R4 + PCA + GPR) workflow. Abbreviations: HC: healthy control, EMCI and LMCI: early and late mild cognitive impairment, AD: Alzheimer's disease.

Time-point	ADNI sample	N	MAE	MSE	Corr (true, pred)	Mean delta	Mean corrected delta (train samples)	Mean corrected delta (ADNI-HC samples)
1	HC	209	6.56	61.19	$r = 0.76$, $p = 4.67e\text{-}40$	−5.97	−5.18	0.00
	EMCI	237	5.76	52.30	$r = 0.72$, $p = 1.07e\text{-}38$	−4.39	−3.78	0.85
	LMCI	127	5.56	46.52	$r = 0.75$, $p = 4.30e\text{-}24$	−3.57	−2.86	2.09
	AD	125	5.18	44.29	$r = 0.66$, $p = 5.00e\text{-}17$	−2.13	−1.20	4.47
2	HC	153	6.56	62.73	$r = 0.73$, $p = 5.46e\text{-}27$	−6.05	−5.27	0.00
	EMCI	197	5.57	50.82	$r = 0.73$, $p = 1.23e\text{-}34$	−4.32	−3.66	1.15
	LMCI	104	5.68	47.75	$r = 0.72$, $p = 6.54e\text{-}18$	−3.25	−2.44	2.88
	AD	61	5.31	44.12	$r = 0.59$, $p = 6.09e\text{-}07$	−0.76	0.31	6.59

Table 6
Pearson's correlation coefficients between corrected brain-age delta using S4_R4 + PCA + GPR workflow and cognitive measures (MMSE, CDR, and FAQ) using age as a covariate from the ADNI sample. The correlations were computed for the whole sample and each diagnostic group (HC, EMCI, LMCI and AD) separately from two timepoints. Abbreviations: MMSE: Mini-Mental State Examination, CDR: Global Clinical Dementia Rating Scale, FAQ: Functional Assessment Questionnaire; HC: healthy control, EMCI and LMCI: early and late mild cognitive impairment, AD: Alzheimer's disease.

	Timepoint-1			Timepoint-2		
	MMSE	CDR	FAQ	MMSE	CDR	FAQ
HC	$N = 68$ $r = -0.202$, $p = 0.101$	$N = 67$ $r = 0.025$, $p = 0.841$	$N = 74$ $r = 0.153$, $p = 0.196$	$N = 153$ $r = -0.065$, $p = 0.427$	$N = 147$ $r = -0.019$, $p = 0.819$	$N = 149$ $r = 0.070$, $p = 0.399$
EMCI	$N = 3$ n.a.	$N = 3$ n.a.	$N = 3$ n.a.	$N = 196$ $r = -0.079$, $p = 0.272$	$N = 194$ **$r = 0.153$, $p = 0.034$**	$N = 193$ $r = 0.091$, $p = 0.211$
LMCI	$N = 2$ n.a.	$N = 2$ n.a.	$N = 2$ n.a.	$N = 103$ **$r = -0.227$, $p = 0.022$**	$N = 102$ $r = 0.115$, $p = 0.253$	$N = 103$ **$r = 0.309$, $p = 0.002$**
AD	$N = 17$ $r = -0.435$, $p = 0.092$	$N = 17$ $r = 0.221$, $p = 0.412$	$N = 26$ $r = 0.244$, $p = 0.240$	$N = 61$ $r = -0.186$, $p = 0.155$	$N = 61$ $r = 0.218$, $p = 0.094$	$N = 61$ **$r = 0.298$, $p = 0.021$**
Whole sample	$N = 90$ **$r = -0.255$, $p = 0.016$**	$N = 89$ $r = 0.114$, $p = 0.290$	$N = 105$ **$r = 0.275$, $p = 0.005$**	$N = 513$ **$r = -0.303$, $p = 2.40e-12$**	$N = 504$ **$r = 0.270$, $p = 7.35e-10$**	$N = 506$ **$r = 0.331$, $p = 2.31e-14$**

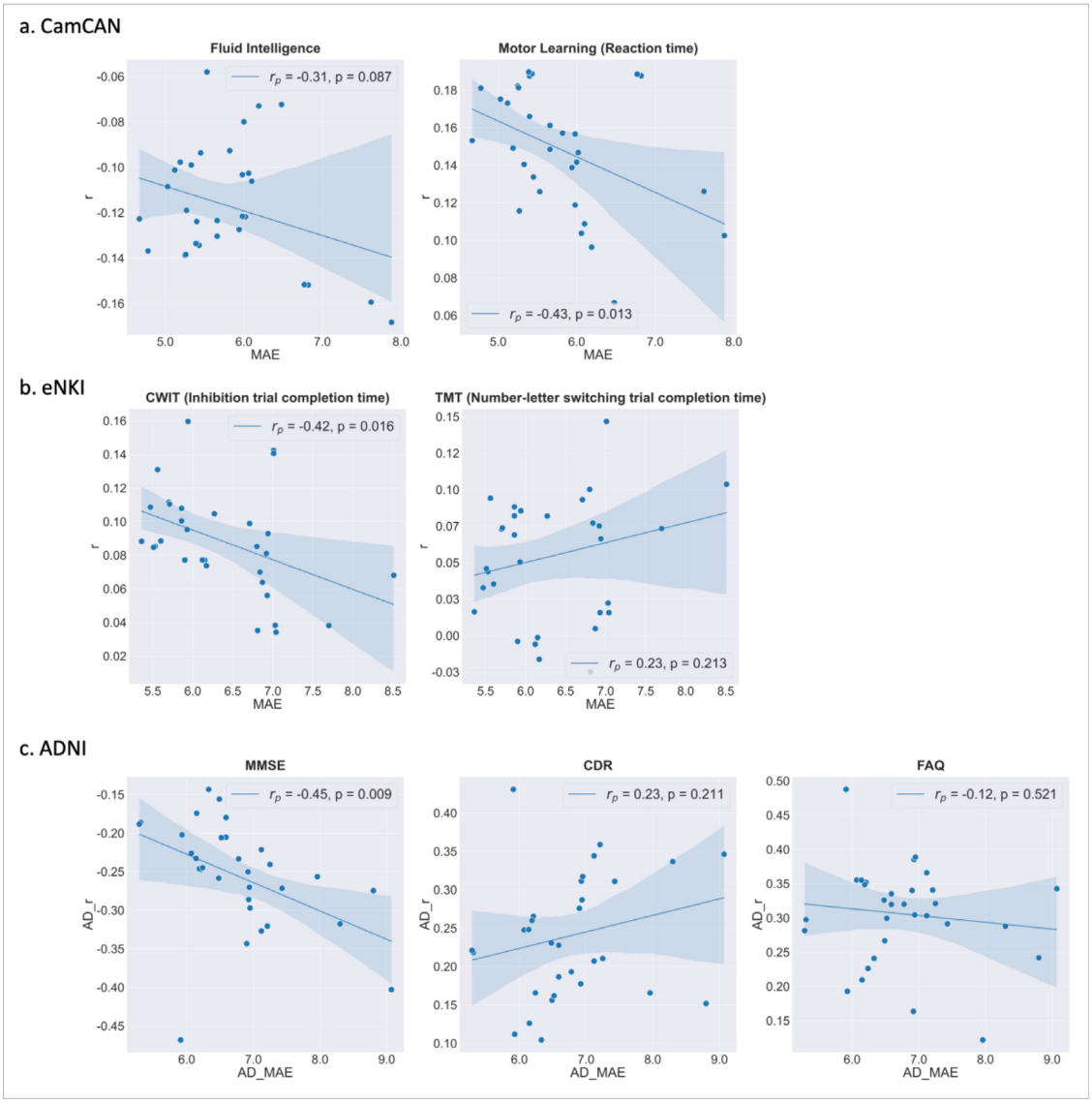
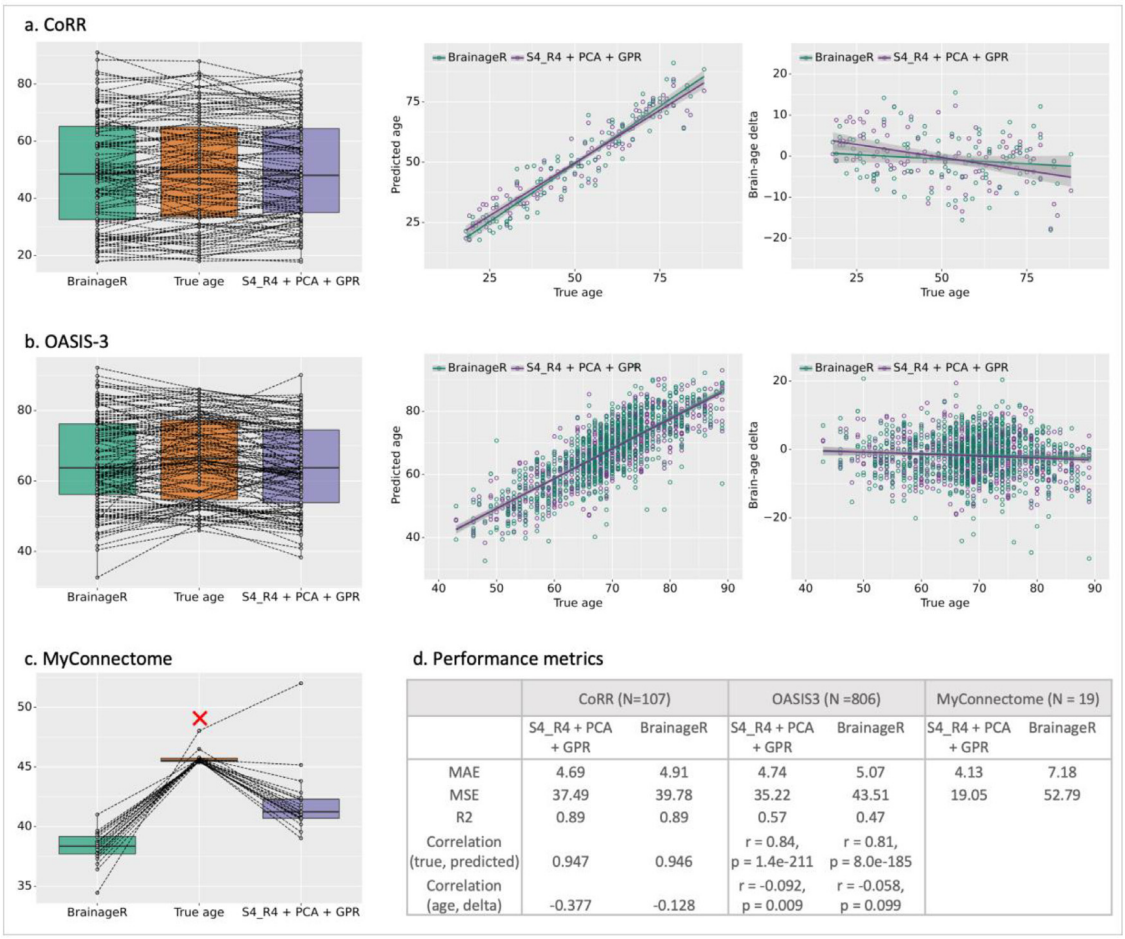


Fig. 5. Correlation between MAE and delta-behavioral correlations obtained using 32 workflows a. CamCAN ($N = 302$) b. eNKI ($N = 340$) c. ADNI ($N = 61$). For CamCAN and eNKI data, the within-dataset delta-behavior correlations with age as a covariate were used. For ADNI data, we used the delta-behavior correlations using corrected delta (corrected using the HC sample) with age as a covariate.



	CoRR (N=107)		OASIS3 (N =806)		MyConnectome (N = 19)	
	S4_R4 + PCA + GPR	BrainageR	S4_R4 + PCA + GPR	BrainageR	S4_R4 + PCA + GPR	BrainageR
MAE	4.69	4.91	4.74	5.07	4.13	7.18
MSE	37.49	39.78	35.22	43.51	19.05	52.79
R2	0.89	0.89	0.57	0.47		
Correlation (true, predicted)	0.947	0.946	r = 0.84, p = 1.4e-211	r = 0.81, p = 8.0e-185		
Correlation (age, delta)	-0.377	-0.128	r = -0.092, p = 0.009	r = -0.058, p = 0.099		

Fig. 6. Comparison of our best workflow (S4_R4 + PCA + GPR) with the brainageR model on a. CoRR dataset (left) the box plot comparing predicted age from two models with true age using a sub-sample of 107 subjects, (center) the scatter plot between the chronological (true) age and predicted age, (right) the scatter plot between the chronological (true) age and brain-age delta. b. OASIS-3 dataset (for visual clarity, the box plot is created using a random sub-sample; $N = 120$) c. MyConnectome dataset (the red cross indicates the outlier scan that was removed from the analysis; final $N = 19$). d. Performance metrics for all datasets. For the CoRR dataset, the table shows average values from 100 iterations of sub-sampled data, but the plots are from one iteration.

3.7. Comparison with brainageR and effect of preprocessing and tissue types

Next, we compared the S4_R4 + PCA + GPR workflow and the brainageR model both trained on the same data using the CoRR, OASIS-3, and MyConnectome datasets (Fig. 6).

In CoRR dataset, S4_R4 + PCA + GPR (mean MAE = 4.69, $r = 0.947$, bias $r = -0.377$) performed better than brainageR (mean MAE = 4.91, $r = 0.946$, bias $r = -0.128$) in MAE (paired t-test: $t = -8.04$, $p = 1.97e\text{-}12$) but brainageR showed a lower mean age bias (Steiger's Z test (Steiger, 1980) $z = -3.31$, $p = 0$; Figs. 6a & S8). There was no significant difference between the mean true and predicted age correlations from two models ($z = 0.133$, p = 0.447).

S4_R4 + PCA + GPR (MAE = 4.74, $r = 0.836$, bias $r = -0.092$) also showed lower MAE than brainageR (MAE = 5.07, $r = 0.805$, bias $r = -0.058$) on the OASIS-3 dataset (Fig. 6b). The predicted ages (paired t-test: $t = -1.37$, $p = 0.17$) and the bias ($z = -1.031$, $p = 0.151$) of the two models were similar but the r value for our model was significantly higher ($z = 3.101$, $p = 0.001$). Test-retest reliability on a sub-sample of the OASIS-3 dataset (retest duration < 3 months) was higher for brainageR (CCC = 0.94 vs. 0.82 for S4_R4 + PCA + GPR). Both models did not show longitudinal consistency at a retest duration of 3–4 years.

Additionally, S4_R4 + PCA + GPR workflow (MAE = 4.13) performed significantly better than brainageR (MAE = 7.18) on the MyConnectome dataset (paired t-test: $t = 9.60$, $p = 1.66e\text{-}08$; Fig. 6c). Note that one outlier scan (true age = 48) was excluded from this analysis (final $N = 19$).

To gain insight into the impact of preprocessing, we compared within-dataset performance of our workflow using SPM preprocessing on IXI and CamCAN datasets. On both datasets, CAT-derived GM features performed better (IXI: MAE = 4.85 years; CamCAN: MAE = 5.01) than SPM-derived GM features (IXI: MAE = 6.25; CamCAN: MAE = 5.82) (Table 7). SPM-derived features from three tissue types performed better (IXI: MAE = 5.08; CamCAN: MAE = 4.88) than using only SPM-derived GM features, indicating that different tissue types carry complementary information (Table 7).

4. Discussion

4.1. Effect of feature space and ML algorithm

The wide range of options available for designing brain-age estimation workflows makes it challenging to disentangle the effect of feature space and ML algorithms. To this end, we investigated 128 workflows constituting combinations of 16 feature representations (voxel-wise and parcel-wise) extracted from GMV images and eight ML algorithms.

Previous studies have shown that the age prediction MAE ranges between ~5–8 years for broad age range data (18–90 years) when using GMV features (Table S1). Our workflows showed performance in a similar range, with some of the workflows generalizing well to data from a

Table 7
Comparison of within-dataset performance between models trained with CAT-preprocessed GM features ($S4\_R4 + PCA + GPR$; our framework), SPM-preprocessed GM features ($S4\_R4_{SPM} + PCA + GPR$) and SPM-preprocessed GM+WM+CSF features ($S4\_R4_{SPM}^{WM+CSF} + PCA + GPR$) on IXI and CamCAN data. Abbreviations: MAE: mean absolute error, MSE: mean squared error, Corr (true, pred): Pearson's correlation between true age and predicted age, Age bias: Pearson's correlation between true age and brain-age delta.

	Workflow	MAE	MSE	Corr (true, pred)	Age bias
IXI (N = 562)	$S4\_R4 + PCA + GPR$	4.85	36.89	r = 0.93, p = 1.03e-247	r = −0.21, p = 7.39e-07
	$S4\_R4_{SPM} + PCA + GPR$	6.25	62.34	r = 0.88, p = 1.15e-181	r = −0.40, p = 1.61e-22
	$S4\_R4_{SPM}^{WM+CSF} + PCA + GPR$	5.08	40.80	r = 0.92, p = 3.98e-234	r = −0.27, p = 1.64e-10
CamCAN (N = 650)	$S4\_R4 + PCA + GPR$	5.01	40.89	r = 0.94, p = 6.45e-307	r = −0.17, p = 1.14e-05
	$S4\_R4_{SPM} + PCA + GPR$	5.82	56.83	r = 0.92, p = 3.87e-258	r = −0.30, p = 2.66e-15
	$S4\_R4_{SPM}^{WM+CSF} + PCA + GPR$	4.88	39.77	r = 0.94, p = 8.29e-308	r = −0.25, p = 1.53e-10

new site. Specifically, the MAE ranged between 4.90–8.48 years in CV and 4.73–8.38 years in test data for within-dataset analysis and for cross-dataset analysis between 4.28–7.39 years and 5.23–8.98 years in CV and test data, respectively. The test MAE and R^2 were highly correlated for both within-dataset and cross-dataset analysis (Tables S2 & S3, Fig. S5). The workflows showed high positive correlations between chronological age and predicted age for within-dataset (r between 0.81–0.93) and cross-dataset (r between 0.82–0.93) analyses. The workflows that performed well in within-dataset analysis also performed well in cross-dataset analysis. The lower cross-dataset CV MAE (4.28–7.39 years) compared to within-dataset CV MAE (4.90–8.48 years) might be because of the larger sample sizes in the cross-dataset analysis or possible overfitting in smaller samples. This corroborates previous studies showing lower errors with larger training sets (Baecker et al., 2021; de Lange et al., 2022), contrary to others that have shown a negative correlation between sample size and CV performance estimates (Wolfers et al., 2015; Varoquaux, 2018). The age range of the training and test data affects the performance estimates. Specifically, when using a narrow age range, performance metrics such as MAE and RMSE are usually better than broad age range evaluations (Cole, 2020; Peng et al., 2021; de Lange et al., 2022). However, the lower errors and hence smaller brain-age delta values in those cases are not necessarily due to better model performance but rather because the predictions are closer to the mean age of the group. Here, our focus was on broad age range models, and the errors we obtained are within the range of what has been previously shown.

Our results showed that the choice of feature space and the ML algorithm both affect the prediction error. In general, feature spaces derived from voxel-wise GMV such as S4_R4, S4_R8, and S0_R4 in combination with GPR, KRR, RVRpoly, and RVRlin algorithms performed well in the within-dataset analysis. The results were similar with PCA retaining 100% variance for some workflows but not all, especially the regularized models (LR and ENR) showed lower performance after PCA (see Supplementary Table S2). This might be because of the different biases of ML algorithms, e.g., due to regularization. It is possible that the sparsity-inducing penalization in addition to PCA leads to lower accuracy models. Some of these selected workflows also performed well on cross-dataset analysis. Specifically, the voxel-wise GMV features smoothed with a 4 mm FWHM kernel and resampled to a spatial resolution of 4 mm, without and with PCA (S4_R4 and S4_R4 + PCA) together with the GPR algorithm performed best in both the within-dataset and cross-dataset analyses. A previous study has reported a voxel size of 3.73 mm^3 and a smoothing kernel of 3.68 mm as the optimal parameters for processing GM images for brain-age prediction with a performance similar to our workflows (Lancaster et al., 2018). In general, parcel-wise features performed worse than voxel-wise features irrespective of the ML algorithm used, suggesting that the GMV summarized from parcels leads to a loss of age-related information. Our results align with a recent study comparing several ML models (GPR-dot product kernel, RVR-linear kernel, and SVR-linear kernel) trained on region-based and voxel-based features with or without PCA on a narrower age range (47–73 years) (Baecker et al., 2021). They found minimal differences in performance due to the ML algorithms with voxel-based features performing better than region-based features.

Our results also indicate that the non-linear algorithm (GPR with RBF kernel) and the kernel-based algorithms (KRR and RVR) outperformed linear algorithms such as RR and LR. Surprisingly, the non-linear RFR algorithm performed the worst irrespective of the feature space used (Fig. S4). This suggests that capturing distributional information using the RBF kernel, as we did using GPR, and use of kernels that capture the similarity between the GMV features in an invariant manner (e.g., Pearson correlation) is beneficial. These results corroborate a recent study that comprehensively evaluated 22 regression algorithms (test MAE between 4.63–7.14 years) in broad age range data (18–94 years) using GMV features and found SVR, KRR, and GPR with a diverse set of kernels to perform well (Beheshti et al., 2022).

In sum, the smoothed and resampled voxel-wise data (such as S4_R4, S4_R8) with either a non-linear or a kernel-based algorithm (GPR with RBF kernel, KRR with polynomial kernel degree (1 or 2), and RVR with linear and polynomial degree 1 kernels) are well suited for brain-age estimation. Sometimes, especially with a large number of features, PCA might help improve performance (Franke et al., 2010; Baecker et al., 2021). However, we found the performance of these workflows with and without PCA to be similar. Therefore, one could use the features directly for immediate interpretability of the models; on the other hand, if computation is a constraint, then the PCA retaining 100% variance could be used without affecting the performance.

Future studies can investigate options to improve model generalizability, such as data harmonization to remove site effects and considerations for population structure (e.g., over-representative of the Caucasian population in the datasets used).

4.2. Test-retest reliability and longitudinal consistency

The brain-age estimates must be reliable within a subject. We found the delta to be reliable over a short scan delay (CoRR: CCC = 0.95–0.98, age range = 20–84; OASIS-3: CCC = 0.76–0.85, age range = 43–80). The reliability of delta within a short scan duration has been reported in previous studies. For example, one study showed an intraclass correlation coefficient (ICC) of 0.96 between deltas from subjects scanned an average of 28.35 ± 1.09 days apart (N = 20, mean age at first scan = 34.05 ± 8.71) (Cole et al., 2017). Another study showed an ICC of 0.93 in young adults from the OASIS-3 dataset (N = 20, age range = 19–34) scanned within a short delay of less than 90 days (Franke and Gaser 2012). Another study found an ICC of 0.81 with a mean interval of 79 days between scans (N = 20, chronological age = 45 years) (Elliott et al., 2021).

Longitudinal consistency, i.e., chronologically proportionate increase in predicted age, is crucial for real-world application. Previous studies have shown lifestyle interventions, such as meditation and exercise (Luders et al., 2016; Steffener et al., 2016), can have positive effects on brain-age, while factors such as smoking and alcohol intake can have adverse effects (Bittner et al., 2021). For instance, 18 months of lifestyle intervention, including diet change and physical activity,

showed attenuated brain-age in a longitudinal sample which correlated with improvement in several physiological measures (Levakov et al., 2022).Thus, lifestyle can lead to different longitudinal brain-age trajectories. However, in our analyses, we assumed that there were no such interventions over the retest duration as the datasets did not provide such information. With this assumption, we expected brain-age to increase proportionally with chronological age.

In support of this assumption, we found a positive linear relationship between the difference in predicted age and the difference in chronological age at a retest duration of 2–3.25 years ($N = 26$; $r = 0.447$, $p = 0.022$) in the CoRR dataset. However, there was no correlation in the OASIS-3 dataset with a retest duration of 3–4 years ($N = 127$; $r = -0.008$, $p = 0.932$). Thus, the evidence of longitudinal consistency was weak. This can be speculatively explained by the maximum test-retest duration of 3–4 years which lies within the range of the MAE for the OASIS-3 dataset (MAE session-1: 5.08 and session-2: 5.86 years, Table S4). Taken together, the high reliability supports the use of brain-age in clinical settings; however, further evaluations are needed to establish longitudinal consistency.

4.3. Effect of bias correction

Most brain-age estimation workflows produce biased results, i.e., overestimation at younger ages and underestimation at older ages (Liang et al., 2019). Therefore, correcting this age bias is important to facilitate individual-level decisions. Here, we adopted a bias correction model that does not use the chronological age of test samples for correction (Cole, 2020), as using chronological age can hamper fair comparison between workflows (de Lange et al., 2022).

The tested workflows generally showed negative associations between chronological age and delta for both within-dataset (r between −0.22 to −0.83) and cross-dataset (r between −0.27 to −0.75) predictions. However, this age bias was less pronounced in more accurate models (Fig. S5). This result is in line with the previous work (de Lange et al., 2022) that showed that if input features are not informative enough to predict age, predictions will be closer to the median or mean age, leading to this bias. Additionally, we found that the data used to estimate the bias correction models can significantly impact the corrected delta. Specifically, within-dataset-derived models corrected the age bias more adequately than cross-dataset models (Fig. S3). This discrepancy might be due to the difference in data properties, e.g., scanner-specific idiosyncrasy, between the training and the out-of-site test data. Our results suggest that a bias correction model might not always work well when applied to a new site, even when the training data itself consists of multiple sites. Consequently, using part of the test data to correct the age bias in the remaining test data works well (as seen in the ADNI data analysis, Section 3.5). However, this might not be feasible when the test sample is small or in the extreme case, a single test subject is available.

How much data is needed for learning a bias correction model is an important but unexplored question. We investigated this by learning bias correction models from sub-samples of the HC subjects from ADNI data. Smaller samples led to higher variance in the efficacy of bias correction models when applied to AD patients (Varoquaux, 2018). For instance, at the smallest sample size ($N = 21$), the average corrected delta of the AD patients varied from 1 to 12 years (Fig. S7, ADNI timepoint 1). It is likely that different studies use different samples for bias correction, so the results should be interpreted and compared with caution. This result shows the importance of using large samples for bias correction and emphasizes careful analysis and reporting of the results.

4.4. Correlation with behavior

Using the selected workflow we observed that the correlation of delta with behavioral measures is sensitive to whether the delta was adjusted for age, either via bias correction or using it as a covariate. For instance, the uncorrected delta was not correlated with FI and motor learning reaction time (in CamCAN data) or CWIT inhibition trial completion time (in eNKI data); however, significant correlations were obtained using age-adjusted delta (Table 4). Thus, it is important to control for age when analyzing correlations between delta and behavioral measures.

Using out-of-sample predictions from within-dataset analysis, we found that a higher uncorrected delta (with age as a covariate) was associated with lower FI, higher motor learning reaction time (from CamCAN data), and lower response inhibition and selective attention, indicated by higher CWIT inhibition trial completion time (from eNKI data). We expected these correlations to be similar to correlations calculated using corrected delta (de Lange and Cole, 2020), as there was no significant age bias. In the CamCAN data, the behavioral correlations using uncorrected delta with age as a covariate and corrected delta were quite similar (FI: $r = -0.154$, $p = 0.0001$ vs. $r = -0.157$, $p = 7.24\text{e-}05$; motor learning reaction time: $r = 0.181$, $p = 0.002$ vs. $r = 0.186$, $p = 0.001$). However, the correlation of CWIT inhibition trial completion time with uncorrected delta with age as a covariate was significant but not when using the corrected delta ($r = 0.109$, $p = 0.045$ vs. $r = 0.094$, $p = 0.084$). This slight difference could potentially be explained by the small effect size and differences inherent in the two methods used for correction.

We also found that there was disagreement between delta-behavior correlations from within-dataset and cross-dataset predictions with age as a covariate. For instance, CamCAN showed significant correlations with FI and motor learning reaction time with within-dataset delta but not with cross-dataset delta. On the other hand, eNKI showed significant correlations only with CWIT inhibition trial completion time using within-dataset delta, but a significant correlation with TMT completion time was found using cross-dataset delta. These results indicate that the subtle differences in predictions can impact behavioral correlations, even though the two predictions were highly correlated (CamCAN: $r = 0.961$, eNKI: $r = 0.962$; Fig. S6). Thus, the delta-behavior correlations, whether using within-dataset or cross-dataset delta, should be interpreted with caution.

Taken together, within-dataset data yields better bias correction models, as we observed in two scenarios, behavioral correlations and delta estimation. However, when enough data are not available, the resulting models may fail to correct the age bias, leading to high variability in the mean delta (Fig. S7). We therefore caution the practitioners and recommend carefully assessing bias correction models, e.g., using bootstrap analysis, before application. We observed that subtle differences in predicted age (within-dataset vs. cross-dataset) lead to different behavioral correlations, which can question the impact of the workflow used for prediction, the analysis method used for computing behavioral correlation (corrected delta versus covariates) and their interaction. Future studies should focus on disentangling such intricacies before applying the brain-age paradigm in practice.

4.5. Higher brain-age delta in neurodegenerative disorders

Neurodegenerative disorders such as AD, MCI, and Parkinson's disease (PD) are accompanied by brain atrophy. Many studies have shown a decrease in global and local GMV in MCI and AD (Good et al., 2001; Karas et al., 2004; Fjell et al., 2014) and also in a broad range of neuropsychiatric disorders (Kaufmann et al., 2019). Consequently, an increased delta, i.e., older appearing brains, has been reported in patients with MCI (3–8 years) and AD (~10 years) (Franke and Gaser 2012; Gaser et al., 2013; Varikuti et al., 2018). We assessed the delta in HC, EMCI, LMCI, and AD patients by applying our best-performing workflow followed by a bias correction model estimated on HC. We found that brain aging is advanced by ~4.5–7 years in AD, ~2–3 years in LMCI, and ~1 year in EMCI (timepoint 1-timepoint 2; Table 5). Furthermore, the delta was correlated with measures associated with disease severity and cognitive impairment in MCI and AD patients. Thus, in line with previous studies, brain-age delta confirmed its potential to indicate accelerated brain aging in neurodegenerative diseases based on structural

MRI data (Franke and Gaser, 2012; Varikuti et al., 2018; Cole et al., 2020; Eickhoff et al., 2021; Lee et al., 2021).

We also show that different workflows can lead to different delta estimates in AD and, consequently, different correlations with cognitive measures (Table S7). In addition, the mean corrected delta in the patient group depends on the type (within-dataset or cross-dataset) and size of sample used for bias correction (Figure S7). Thus, the results should be interpreted with caution when comparing different studies.

4.6. Relationship of MAE with delta and delta-behavior correlations

The utility of age prediction models lies in their application to capture atypical aging. However, to achieve this, it is imperative to minimize the methodological variance, due to decisions in feature space and ML algorithms, by building accurate models so that the resulting brain-age delta captures biological variance. A recent study has shown that delta from overfitted models (i.e., with higher training accuracy) results in smaller differences in AD vs. CN, while delta from a model with comparatively lower (training) accuracy captures biological variance (Bashyam et al., 2020). However, our analyses and model selection was based on nested cross-validation. Therefore, our accurate models cannot be considered overfitted.

In healthy samples, higher accuracy (lower MAE) was associated with higher delta-motor learning reaction time (CamCAN) and delta-CWIT inhibition trial completion time (eNKI) associations. In contrast, in AD patients, models with lower accuracy (higher MAE) showed a stronger delta-MMSE correlation. This observation that some less accurate models can better capture the delta-behavioral correlation better in AD is in line with a previous study (Bashyam et al., 2020) (Fig. 5 and Table S7). These contrasting observations in healthy and patient cohorts make it difficult to develop a model selection strategy based on delta-behavioral correlations.

The corrected mean delta in AD (corrected using the CN sample, indicative of separation between CN and AD), for the 32 workflows ranged from 5.43 to 10.01 years. Some moderately accurate models, e.g., S0_R4 + LR (delta = 7.27, MAE = 5.91 years), showed a high delta for AD and a strong correlation with AD scales (Table S7). However, the model with the highest delta (173 + RFR: delta = 10.01, MAE: 9.07 years) showed a comparatively weaker correlation with behavior. Moreover, similarly performing models (S0_R4 + LR: delta = 7.27, MAE = 5.91 years vs. S8_R4 + KRR: delta = 7.17, MAE = 6.59 years) showed quite different correlation with behavior. This indicates a non-linear relationship between the models' MAEs, deltas, and behavioral correlations.

Based on these results, we speculate that perhaps using adequately regularized models in the patient population can be beneficial even if they show a lower accuracy. It might be possible that regularization pushes the models to focus on fewer specific features containing typical aging-related signal. This in turn could lead to lower accuracy models (as it downweighs some features) but also leads to delta estimates that are more informative of atypical aging.

Taken together, comparing models based on their performance on patient data and delta-behavior correlations is a promising but open topic. In particular, it is unclear which delta-behavioral correlation to use, and generalizability of models across behavioral scores, samples, and disorders remains unknown. Further studies are needed to define appropriate procedures for model selection based on such criteria.

4.7. Comparison with brainageR and effect of preprocessing and tissue types

Using the same training data as brainageR, our workflow outperformed brainageR in terms of MAE in three datasets; CoRR (N = 107; mean MAE = 4.69 vs. 4.91), OASIS-3 (N = 806; MAE = 4.74 vs. 5.07), and MyConnectome (N = 19; MAE = 4.13 vs. 7.18). However, the bias of our model was similar or higher than that of brainageR and its test-retest reliability was lower (OASIS-3, N = 36; CCC = 0.82 vs. CCC = 0.94). Overall, our workflow showed lower MAE, higher correlation between true and predicted age but also higher age bias compared to brainageR. These differences are likely driven by differences in preprocessing, and the use of three tissue types by brainageR as opposed to us using only GM. To investigate this further, we performed two additional analyses.

Different VBM tools can provide different GMV estimates, influencing the estimated association with age (Tavares et al., 2019; Antonopoulos et al., 2023). The CAT-derived GMV features performed better than SPM preprocessing (both with S4_R4 + PCA for feature extraction together with the GPR algorithm for learning) in terms of MAE (e.g., IXI: MAE = 4.85 vs. 6.25), the correlation between true and predicted age (r = 0.93 vs. 0.88, p < 1e-6 both) and age bias (r = −0.21 vs. r = −0.40, p < 1e-6 both) (Table 7). We further found that the predictions when using three tissue types from SPM (GM, WM, and CSF) were better (IXI: MAE = 5.08, r = 0.92, p < 1e-6, bias: r = −0.27, p < 1e-6). This is in line with a previous study that showed a slight performance improvement when using both GM and WM compared to only GM (Cole et al., 2017). Features from different tissue types may carry complementary information regarding age, providing better predictions and lower age bias. Many previous studies have used GM and WM together as features (Franke and Gaser, 2012; Cole et al., 2017; Cole et al., 2018, 2020), and others have used all three tissue types (Monté-Rubio et al., 2018; Xifra-Porxas et al., 2021; Hobday et al., 2022). CAT-derived GMV performed similarly to SPM-derived three tissue types with slightly lower age bias for the former (Table 7), showing the suitability of GM for this task following its clinical relevance in neurodegenerative disorders (Karas et al., 2004; Wu et al., 2021). Further studies are needed to cleanly disentangle the effect of tissue types on different performance criteria investigated here.

5. Conclusion

Numerous choices exist for designing a workflow for age prediction. The systematic evaluation of different workflows on the same data in different scenarios (within-dataset, cross-dataset, test-retest reliability, and longitudinal consistency) revealed a substantial impact of feature representation and ML algorithm choices. Notably, voxel-wise GM features, especially smoothed with a 4 mm FWHM kernel and resampled to a spatial resolution of 4 mm (S4_R4), were better than parcel-wise features. Additionally, performing PCA did not affect the prediction performance, but it can help reduce computational resources. ML algorithms, including Gaussian process regression with the radial basis kernel, kernel ridge regression with polynomial kernel degree 1 or 2, and relevance vector machine with linear and polynomial degree 1 kernels, performed well. Overall, some workflows performed well on out-of-site data and showed high test-retest reliability but only moderate longitudinal reliability. Consistent with the literature, we found a higher delta in Alzheimer's and mild cognitive impairment patients after correcting the delta with a large sample of controls. Our results provide evidence for the potential future application of delta as a biomarker but also caution regarding analysis setup and data used for behavioral correlations and bias correction. Findings from the current study can serve as guidelines for future brain-age prediction studies.

Ethics statement

Ethical approval and informed consent were obtained locally for each study covering both participation and subsequent data sharing. The ethics proposals for the use and retrospective analyses of the datasets were approved by the Ethics Committee of the Medical Faculty at the Heinrich-Heine-University Düsseldorf.

Declaration of Competing Interest

The authors report no competing interests

Credit authorship contribution statement

Shammi More: Formal analysis, Software, Validation, Visualization, Writing – original draft. **Georgios Antonopoulos:** Data curation, Writing – review & editing. **Felix Hoffstaedter:** Data curation, Writing – review & editing. **Julian Caspers:** Supervision, Writing – review & editing. **Simon B. Eickhoff:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Kaustubh R. Patil:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Data availability

The codes used for preprocessing, feature extraction, model training and prediction are available at https://github.com/juaml/brainage_estimation. All the datasets used except the 1000BRAINS data are available publicly (might require registration and approval).

Acknowledgments

This study is supported by Deutsche Forschungsgemeinschaft (DFG, PA 3634/1-1 and EI 816/21-1), the National Institute of Mental Health, the Helmholtz Portfolio Theme “Supercomputing and Modelling for the Human Brain” and the European Union’s Horizon 2020 Research and Innovation Program grant agreement 945539 (HBP SGA3).

Data for the MyConnectome project were obtained from the OpenNeuro database (ds000031).

The clinical data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of the ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see www.adni-info.org. Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.119947.

References

Antonopoulos, G., More, S., Raimondo, F., Eickhoff, S.B., Hoffstaedter, F. and Patil, K.R. 2023. A systematic comparison of VBM pipelines and their application to age prediction. *BioRxiv*.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38 (1), 95–113.

Ashburner, J., Friston, K.J., 2011. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. Neuroimage 55 (3), 954–967.

Baecker, L., Dafflon, J., da Costa, P.F., et al., 2021. Brain age prediction: a comparison between machine learning models using region- and voxel-based morphometric data. Hum. Brain Mapp. 42 (8), 2332–2346.

Bashyam, V.M., Erus, G., Doshi, J., et al., 2020. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. Brain J. Neurol. 143 (7), 2312–2324.

Beheshti, I., Ganaie, M.A., Paliwal, V., Rastogi, A., Razzak, I., Tanveer, M., 2022. Predicting brain age using machine learning algorithms: a comprehensive evaluation. IEEE J. Biomed. Health Inform. 26 (4), 1432–1440.

Beheshti, I., Nugent, S., Potvin, O., Duchesne, S., 2019. Bias-adjustment in neuroimaging-based brain age frameworks: a robust scheme. Neuroimage Clin. 24, 102063.

Bittner, N., Jockwitz, C., Franke, K., et al., 2021. When your brain looks older than expected: combined lifestyle risk and BrainAGE. Brain Struct. Funct. 226 (3), 621–645.

Boyle, R., Jollans, L., Rueda-Delgado, L.M., et al., 2021. Brain-predicted age difference score is related to specific cognitive functions: a multi-site replication analysis. Brain Imaging Behav. 15 (1), 327–345.

Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Yeo, B.T.T., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. J. Neurophysiol. 106 (5), 2322–2345.

Caspers, S., Moebus, S., Lux, S., et al., 2014. Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. Front. Aging Neurosci. 6, 149.

Chen, J., Liu, J., Calhoun, V.D., et al., 2014. Exploration of scanning effects in multi-site structural MRI studies. J. Neurosci. Methods 230, 37–50.

Cole, J.H., 2020. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. Neurobiol. Aging 92, 34–42.

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. Ann. Neurol. 77 (4), 571–581.

Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., et al., 2017a. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. Neuroimage 163, 115–124.

Cole, J.H., Raffel, J., Friede, T., et al., 2020. Longitudinal assessment of multiple sclerosis with the brain-age paradigm. Ann. Neurol. 88 (1), 93–105.

Cole, J.H., Ritchie, S.J., Bastin, M.E., et al., 2018. Brain age predicts mortality. Mol. Psychiatry 23 (5), 1385–1392.

Cole, J.H., Underwood, J., Caan, M.W.A., et al., 2017b. Increased brain-predicted aging in treated HIV disease. Neurology 88 (14), 1349–1357.

Eickhoff, C.R., Hoffstaedter, F., Caspers, J., et al., 2021. Advanced brain ageing in Parkinson’s disease is related to disease duration and individual impairment. Brain Commun. 3 (3), fcab191.

Elliott, M.L., Belsky, D.W., Knodt, A.R., et al., 2021. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. Mol. Psychiatry 26 (8), 3829–3838.

Fan, L., Li, H., Zhuo, J., et al., 2016. The human brainnetome atlas: a new brain atlas based on connectional architecture. Cereb. Cortex 26 (8), 3508–3526.

Fjell, A.M., McEvoy, L., Holland, D., Dale, A.M., Walhovd, K.B.Alzheimer’s Disease Neuroimaging Initiative, 2014. What is normal in normal aging? Effects of aging, amyloid and Alzheimer’s disease on the cerebral cortex and the hippocampus. Prog. Neurobiol. 117, 20–40.

Franke, K., Gaser, C., 2012. Longitudinal changes in individual *BrainAGE* in healthy aging, mild cognitive impairment, and Alzheimer’s disease. GeroPsych 25 (4), 235–245 *(Bern)*.

Franke, K., Gaser, C., Manor, B., Novak, V., 2013. Advanced BrainAGE in older adults with type 2 diabetes mellitus. Front. Aging Neurosci. 5, 90.

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 (3), 883–892.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.

Galluzzi, S., Beltramello, A., Filippi, M., Frisoni, G.B., 2008. Aging. Neurolog. Sci. 29 (Suppl 3), 296–300.

Gaser, C., Dahnke, R., Thompson, P.M., Kurth, F., Luders, E. and Alzheimer’s Disease Neuroimaging Initiative 2022. CAT – a computational anatomy toolbox for the analysis of structural MRI data. *Biorxiv*.

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 (6), e67346.

Giorgio, A., Santelli, L., Tomassini, V., et al., 2010. Age-related changes in grey and white matter structure throughout adulthood. Neuroimage 51 (3), 943–951.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 14 (1 Pt 1), 21–36.

Grinsztajn, L., Oyallon, E. and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on tabular data? *arXiv*.

Gutierrez Becker, B., Klein, T., Wachinger, C.Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing, 2018. Gaussian process uncertainty in age estimation as a measure of brain abnormality. Neuroimage 175, 246–258.
Hahn, T., Ernsting, J., Winter, N.R., et al., 2022. An uncertainty-aware, shareable, and transparent neural network architecture for brain-age modeling. Sci. Adv. 8 (1), eabg9471.
He, T., Kong, R., Holmes, A.J., et al., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage 206, 116276.
Hobday, H., Cole, J.H., Stanyard, R.A., et al., 2022. Tissue volume estimation and age prediction using rapid structural brain scans. Sci. Rep. 12 (1), 12005.
Jack, C.R., Bernstein, M.A., Fox, N.C., et al., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.
Jiang, H., Lu, N., Chen, K., et al., 2019. Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks. Front. Neurol. 10, 1346.
Jolliffe, I.T., 2002. Principal Component Analysis, 2nd ed. Springer-Verlag, New York.
Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., et al. 2019. Deep learning based brain age prediction uncovers associated sequence variants. *Biorxiv*.
Jovicich, J., Czanner, S., Greve, D., et al., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 30 (2), 436–443.
Karas, G.B., Scheltens, P., Rombouts, S.A.R.B., et al., 2004. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. Neuroimage 23 (2), 708–716.
Kaufmann, T., van der Meer, D., Doan, N.T., et al., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat. Neurosci. 22 (10), 1617–1623.
LaMontagne, P.J., Benzinger, T.L.S., Morris, J.C., et al. 2019. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*.
Lancaster, J., Lorenz, R., Leech, R., Cole, J.H., 2018. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. Front. Aging. Neurosci. 10, 28.
de Lange, A.-M.G., Anatürk, M., Rokicki, J., et al., 2022. Mind the gap: performance metric evaluation in brain-age prediction. Hum. Brain Mapp..
de Lange, A.-M.G., Cole, J.H., 2020. Commentary: correction procedures in brain-age prediction. Neuroimage Clin. 26, 102229.
de Lange, A.-M.G., Kaufmann, T., van der Meer, D., et al., 2019. Population-based neuroimaging reveals traces of childbirth in the maternal brain. Proc. Natl. Acad. Sci. U.S.A. 116 (44), 22341–22346.
Le, T.T., Kuplicki, R.T., McKinney, B.A., et al., 2018. A nonlinear simulation framework supports adjusting for age when analyzing brainage. Front. Aging. Neurosci. 10, 317.
Lee, W.H., Antoniades, M., Schnack, H.G., Kahn, R.S., Frangou, S., 2021. Brain age prediction in schizophrenia: does the choice of machine learning algorithm matter? *Psychiatry research*. Neuroimaging 310, 111270.
Levakov, G., Kaplan, A., Meir, A.Y., et al. 2022. The effect of 18 months lifestyle intervention on brain age assessed with resting-state functional connectivity. *medRxiv*.
Liang, H., Zhang, F., Niu, X., 2019. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. Hum. Brain Mapp. 40 (11), 3143–3152.
Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255–268.
Löwe, L.C., Gaser, C., Franke, K.Alzheimer's Disease Neuroimaging Initiative, 2016. The effect of the APOE genotype on individual brainage in normal aging, mild cognitive impairment, and Alzheimer's disease. PLoS One 11 (7), e0157514.
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.
Monté-Rubio, G.C., Falcón, C., Pomarol-Clotet, E., Ashburner, J., 2018. A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. Neuroimage 178, 753–768.
More, S., Eickhoff, S.B., Caspers, J., Patil, K.R, 2021. Confound removal and normalization in practice: a neuroimaging based sex prediction case study. In: Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., Van Hoecke, S. (Eds.), *ECML PKDD 2020: Demo Track*. Lecture Notes in Computer Science. Springer International Publishing, Ghent, Belgium, pp. 3–18.
Nooner, K.B., Colcombe, S.J., Tobe, R.H., et al., 2012. The NKI-rockland sample: a model for accelerating the pace of discovery science in psychiatry. Front. Neurosci. 6, 152.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res..
Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. 68, 101871.
Petersen, R.C., Aisen, P.S., Beckett, L.A., et al., 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology 74 (3), 201–209.
Poldrack, R.A., Laumann, T.O., Koyejo, O., et al., 2015. Long-term neural and physiological phenotyping of a single human. Nat. Commun. 6, 8885.
Richard, G., Kolskår, K., Sanders, A.-.M., et al., 2018. Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry. PeerJ 6, e5908.
Schaefer, A., Kong, R., Gordon, E.M., et al., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 28 (9), 3095–3114.
Schulz, M.-.A., Yeo, B.T.T., Vogelstein, J.T., et al., 2020. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. Nat. Commun. 11 (1), 4238.
Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., Miller, K.L., 2019. Estimation of brain age delta from brain imaging. Neuroimage 200, 528–539.
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. Neurobiol. Aging 40, 138–144.
Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. 87 (2), 245–251.
Su, L., Wang, L., Hu, D., 2013. Predicting the age of healthy adults from structural MRI by sparse representation. In: Yang, J., Fang, F., Sun, C (Eds.), *Intelligent Science and Intelligent Data Engineering*. Lecture notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 271–279.
Tavares, V., Prata, D., Ferreira, H.A., 2019. Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study. J. Neurosci. Methods 334, 108565.
Taylor, J.R., Williams, N., Cusack, R., et al., 2017. The Cambridge centre for ageing and neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage 144 (Pt B), 262–269.
Thompson, N.C., Greenewald, K., Lee, K. and Manso, G.F. 2020. The computational limits of deep learning. *arXiv*.
Treder, M.S., Shock, J.P., Stein, D.J., du Plessis, S., Seedat, S., Tsvetanov, K.A., 2021. Correlation constraints for regression models: controlling bias in brain age prediction. Front. Psychiatry 12, 615754.
Varikuti, D.P., Genon, S., Sotiras, A., et al., 2018. Evaluation of non-negative matrix factorization of grey matter in age prediction. Neuroimage 173, 394–410.
Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180 (Pt A), 68–77.
Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145 (Pt B), 166–179.
Vidal-Pineiro, D., Wang, Y., Krogsrud, S.K., et al., 2021. Individual variations in "brain age" relate to early-life factors more than to longitudinal brain change. Elife 10.
Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci. Biobehav. Rev. 57, 328–349.
Wu, Z., Peng, Y., Hong, M., Zhang, Y., 2021. Gray matter deterioration pattern during Alzheimer's disease progression: a regions-of-interest based surface morphometry study. Front. Aging Neurosci. 13, 593898.
van Wynsberghe, A., 2021. Sustainable AI: AI for sustainability and the sustainability of AI. AI Ethics.
Xifra-Porxas, A., Ghosh, A., Mitsis, G.D., Boudrias, M.-.H., 2021. Estimating brain age from structural MRI and MEG data: insights from dimensionality reduction techniques. Neuroimage 231, 117822.
Zhao, Y., Klein, A., Castellanos, F.X., Milham, M.P., 2019. Brain age prediction: cortical and subcortical shape covariation in the developing human brain. Neuroimage 202, 116149.
Zuo, X.-.N., Anderson, J.S., Bellec, P., et al., 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data 1, 140049.