BrainAGE: Revisited and reframed machine learning workflow

Author names and affiliations:

Polona Kalc[1], Robert Dahnke[1,2], Felix Hoffstaedter[4,5], Christian Gaser[1,2,3], and Alzheimer's Disease Neuroimaging Initiative

[1]Structural Brain Mapping Group, Department of Neurology, Jena University Hospital

[2]Department of Psychiatry and Psychotherapy, Jena University Hospital

[3]German Center for Mental Health (DZPG)

[4]Forschungszentrum Jülich, Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7)

[5]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf

Corresponding author: Christian Gaser

Address: Am Klinikum 1, 07747 Jena - DE

Abstract

Since the introduction of the BrainAGE method (Franke et al., 2010), novel machine learning methods of brain age prediction have continued to emerge. The idea of estimating the chronological age from magnetic resonance images proved to be an interesting field of research due to the relative simplicity of its interpretation and its potential use as a biomarker of brain health. We revised the BrainAGE approach, originally utilising relevance vector regression (RVR), and substituted it with Gaussian process regression (GPR), which enables more stable processing of larger datasets, such as the UK Biobank (UKB). We tested the performance of the new model under several different conditions and investigated its validity on healthy and neurologically impaired subjects of the ADNI sample and on a synthetic dataset of neocortical thinning. The results show an improved performance of the reframed model with a MAE of 1.97 years on a UKB sample and a significant difference in BrainAGE between healthy participants and patients with Alzheimer's disease. Moreover, the workings of the algorithm show meaningful effects for a simulated neocortical atrophy dataset. The results demonstrate that the new improved algorithm provides reliable and valid brain age estimations.

Key words

brain age, UK Biobank, pre-processing, machine learning, mean absolute error, structural MRI, Gaussian process regression

Data availability statement

The data used for this work was obtained from the UK Biobank Resource (Project Number 41655). Due to the nature of the data sharing agreement, we are not allowed to publish the data.

ADNI dataset is publicly available at https://ida.loni.usc.edu/login.jsp?project=ADNI and IXI at https://brain-development.org/ixi-dataset/. The synthetic dataset from Rusak et al. (2022) is available at the CSIRO data access portal.

Conflict of interest

The authors declare no conflict of interest.

Introduction

Our original BrainAGE framework (Franke et al., 2010) is a supervised machine learning method that employs relevance vector regression (RVR) in order to predict the chronological age of a subject based on a single T1-weighted magnetic resonance (MR) scan of its brain. The difference between the apparent age of the subject's brain as predicted by the algorithm and its chronological age is the so-called brain age gap estimation (BrainAGE). A positive gap implies more prominent structural brain changes that commonly occur with ageing progression, e.g. decrease of grey matter volume (Bethlehem et al., 2022), whereas a negative difference signifies a slower ageing in comparison to other subjects within a sample. The BrainAGE method was introduced more than a decade ago and was successfully applied to various subject groups, such as adolescents (Franke et al., 2012), elderly people with mild cognitive impairment (MCI) and Alzheimer's disease (AD; Franke et al., 2010; Franke & Gaser, 2012), musicians (Rogenmoser et al., 2018), meditators (Luders et al., 2016), and even other animal species, such as rodents and non-human primates (Franke et al., 2017). The method proved to be a reliable biomarker of AD (Franke & Gaser, 2012), Parkinson's disease (Eickhoff et al. 2021), and is associated with brain alterations in schizophrenia (Nenadić et al., 2017) and type 2 diabetes (Franke et al., 2013). A complete review can be found in Franke & Gaser (2019) or Cole & Franke (2017).

Since the introduction of the concept of brain age as a biomarker of ageing brains, numerous machine learning and deep learning models of estimating the age of the brain based on different MR imaging modalities have emerged (e.g., Beheshti et al., 2020; Cole et al., 2018; Hahn et al. 2022; Smith et al., 2019; Varzandian et al., 2021). The advancement of deep learning methods in the field of brain age prediction (e.g., Abrol et al., 2021; Gong et al., 2020; Hahn et al., 2022; Lombardi et al., 2021; Peng et al., 2021) and the availability of large multimodal imaging data collection, such as the UK Biobank (UKB; https://www.ukbiobank.ac.uk/), pushed the performance of the brain age prediction to unprecedented levels, managing to decrease mean absolute errors (MAE; i.e., a metric commonly used to compare the accuracy of prediction) to as low as 2.14 years (Peng et al., 2021).

The aim of the study was to revise the original BrainAGE machine learning framework to be able to keep pace with the advancement of deep learning algorithms in dealing with large datasets, while retaining the parsimonious and straightforward nature of voxel-based morphometry (VBM). Studies testing the performance of multiple algorithms in brain age estimation showed a good performance of Gaussian process regression (GPR) (Baecker et al., 2021; Beheshti et al., 2022; Cole et al., 2015; Cole et al., 2018; More et al. 2023), which can be seen as an alternative approach to artificial neural networks (Chen & Ren 2009). We therefore implemented GPR in the new BrainAGE framework and utilised model ensembling to limit overfitting and improve prediction accuracy (van Veen et al., 2015). We replaced the initial BrainAGE framework (Franke et al., 2010; Franke & Gaser, 2012) with a combination of models using grey- or white matter tissue segments (of varying spatial resolutions and smoothing sizes) associated by (weighted) averaging and stacking.

In order to evaluate the performance of our novel BrainAGE algorithm's and to allow for straightforward comparison to results of other groups, we focused on the UK Biobank (UKB) cohort. The UKB sample is a generally available, very large biomedical database and research resource that contains genetic, lifestyle and health information from half a million UK Biobank participants. We investigated the effects of various conditions (e.g., dimension reduction, training sample size) and examined the reliability of the predictions. Furthermore, we tested the out-of-sample performance of the new model trained on the UKB participants on a sample of subjects ranging from healthy participants to AD patients from the ADNI cohort. Finally, we applied it to a novel synthetic dataset of global neocortical thinning (Rusak et al. 2002) to test how the BrainAGE algorithm performs on a simulated dataset and how neocortical thinning influences the brain age prediction.

Method

Datasets

This research has been conducted using data from UK Biobank under the application number 41655. The UK Biobank has ethical approval from the North West Multi-Centre Research Ethics Committee (MREC) and is in possession of informed consents from the study cohort participants.

Participants with pre-existing neurological or psychiatric diagnoses were excluded from the analyses (UKB data-fields #41202-0.0 to 41202-0.78 Diagnoses - main ICD10: presence of F0-F99 Mental and behavioural disorders). Overall, we analysed 36840 T1-weighted MR scans (52% women; age range: 45–82; M_{Age} = 64.09, SD_{Age} = 7.53 years). Some of the analyses were run on a pseudo random subsample of 7241 scans (52% women; age range: 45–82; M_{Age} = 64.08, SD_{Age} = 7.54 years), including only subject IDs starting with 1* to avoid the introduction of selection bias. For the cross-site prediction, we randomly sampled a fifth of the full sample (n = 7368) to obtain two size-matched subsamples from the scanning site 1 (Cheadle) and the scanning site 3 (Newcastle).

Additionally, 695 T1w images from Alzheimer's Disease Neuroimaging Initiative (ADNI, https://adni.loni.usc.edu/) were used to validate the model on a sample including healthy

subjects (HC, n = 231, M_{Age} = 76.01±5.01), subjects with stable MCI (sMCI, n = 100, M_{Age} = 75.40±7.27), subjects who progressed from MCI to Alzheimer's dementia (pMCI, n = 164, M_{Age} = 74.57±7.03), and AD patients (n = 200, M_{Age} = 75.64±7.72).

To further validate the model, we used a synthetic global neocortical atrophy dataset by Rusak et al. (2022), which was derived from ADNI baseline scans of 20 subjects without AD (n = 20; M_{Age} = 70.65±5.39) and used to simulate neocortical thinning, progressing from 0 to 0.1 mm or 1 mm thickness loss, with steps of 0.01 mm or 0.1 mm between the time points, respectively, resulting in total of 400 synthetic images.

Additionally, the IXI sample was used for testing the adjusted BrainAGE for training sample characteristics. An overview of the datasets used in this study can be seen in Table 1.

Table 1. *Characteristics of the datasets used in this study*.

Dataset		*N* (sessions)	Age range	M ± SD	Males/females
UKB		36840	45–82	64.09±7.53	17557/19283
ADNI	HC	231	60–90	76.01±5.01	119/112
	sMCI	100	58–88	75.40±7.27	66/34
	pMCI	164	55–88	74.57±7.03	97/67
	AD	200	55–91	75.64±7.72	103/97
IXI		547	19–86	48.12±16.61	242/305
Synthetic dataset		20 (20)	59–78	70.65±5.39	10/10

Data acquisition

The information regarding data acquisition for the UKB dataset is available in the UK Biobank Imaging documentation (https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf). In short, T1-weighted MR images were acquired on a Siemens Skyra 3T scanner with a 32-channel RF receive head coil. The MPRAGE sequence was used with 1-mm isotropic resolution, inversion/repetition time = 880/2000 ms, acquisition time: 5 minutes, FOV: 208x256x256 matrix, in-plane acceleration factor = 2.

In the ADNI sample used, T1w images were acquired on 1.5T scanners manufactured by Siemens, Philips, and GE with single-channel RF coils. The MPRAGE sequences typically had about 0.95 mm in-plane resolution and 1.2 mm slice-thickness, inversion/repetition time = 853-1000/2300-3000ms, acquisition time: 7:11–9:38 minutes, FOV: 240-260x240x160-208, no acceleration (Jack et al., 2008; see also https://adni.loni.usc.edu/methods/mri-tool/mri-acquisition/).

The IXI sample comprises data from 3 different scanners (Philips 1.5T and 3T, GE 1.5T) with about 0.95 mm in-plane resolution and 1.2 mm slice-thickness (https://brain-development.org/ixi-dataset).

Data preprocessing

T1-weighted images were segmented into grey matter (GM) and white matter (WM) and (affinely/non-linearly) normalised by using default preprocessing of the CAT12.7 Toolbox (Gaser et al., 2022) as standalone version compiled under Matlab 2019b and SPM12 (Wellcome Centre for Human Neuroimaging, https://www.fil.ion.ucl.ac.uk/spm/) massively parallelized on the JURECA High Performance Computing system (JSC). The preprocessing utilises the unified segmentation (Ashburner & Friston, 2005) to remove B0-inhomogeneities and create an initial segmentation that is used for (local) intensity scaling and adaptive non-local means denoising (Manjon et al., 2010). An adaptive maximum a posteriori (AMAP; Rajapakse et al., 1997) segmentation with a hidden Markov random field (Cuadra et al., 2005) and partial volume effect model (Tohka et al., 2004) is utilised to create the final segmentation. For the optional non-linear registration, the Shooting method (Ashburner & Friston, 2011) with modulation was used.

The synthetic dataset was processed with the CAT12.8 version r2443, which was optimised to handle the atypical non-zero skull-stripped backgrounds, and therefore enabled more stable processing of the data. All other relevant parameters were held constant.

BrainAGE models estimation

In order to keep the process as parsimonious as possible, linear affine registration of tissue segments to the MNI152Nlin2009cAsym template was used as default, except for a single non-linearly registered GM test case. Eight combinations of single tissue class models (GM/WM) with varying parameterisations of spatial resolution (4 mm / 8 mm) and Gaussian smoothing (FWHM: 4 mm / 8 mm) were trained and processed in the same way. With the exception of the first testing condition (see the test conditions below), principal component analysis (PCA) using singular value decomposition was applied to all the models while using n–1 principal components, where n is the minimum of voxel number or sample size. We employed a GPR model instead of the previously used RVR (see Figure 1A for the comparison of performance of both algorithms). This GPR uses a linear covariance function, a constant mean function and a Gaussian likelihood function with the hyperparameters set to 100 for the constant mean function and –1 for the likelihood function (Rasmussen & Williams, 2006). A conjugate gradient method was used for numerical optimization in the GPR.

A 10-fold cross-validation was performed with stratified chronological age to preserve the age distribution in each fold. The estimated brain age was age-bias adjusted by a linear term as described in Smith et al. (2019). Single tissue class models were ensembled by averaging (unless stated otherwise).

The estimations were run in Matlab 2021a (Mathworks Inc., Natick, MA, USA).

Results

With respect to the change in the algorithm, we tested the novel machine learning framework performance under different conditions outlined in the following section, jointly reporting methods and results. We show the MAEs of the average-ensembled age-bias corrected models and Pearson's correlation coefficients between the predicted age and the chronological age. Please see the Supplementary material (Tables S2 & S3) for the performance of all tested models without age-bias correction.

I. The effect of principal component analysis

In this step, we tested the effect of implementing PCA using singular value decomposition, which was applied to the training data only and the resulting transformation was used on the test sample to avoid data leakage. The performance of the model greatly improved with the drop from MAE of 3.13 years (r_{Age} = 0.89) to MAE of 2.46 years (r_{Age} = 0.93). All subsequent estimations on the full UKB dataset were therefore run on reduced data.

II. Tissue class combination by model ensembling

Single tissue class models using only one linearly registered brain tissue class (either GM, WM, or non-linearly registered modulated GM), with varying parameterisations of spatial resolution (4 mm / 8 mm) and Gaussian smoothing (FWHM: 4 mm / 8 mm) were ensembled in order to minimise the MAE of the brain age prediction (Figure 1B). Three different ensembling approaches were tested, namely (i) their average performance, (ii) the weighted mean, and (iii) GPR stacking of the estimated models. The weighted mean was calculated with weighting w.r.t. squared MAE. The GPR stacking was done by utilising predictions from the first level models as features for the second GPR learner.

We additionally tested the performance of mean-ensembled single tissue class models (either GM-, non-linearly registered modulated GM-, WM-, and concatenated GM and WM vectors).

The results of ensembling different brain tissue classes are presented in Figure 1A and 1B for the subsample and the full sample, respectively. Model ensembling provides a more accurate BrainAGE estimation, with the most striking improvement obtained by stacking. Using the full UKB sample, GPR stacking resulted in 2.18 years of MAE for uncorrected BrainAGE prediction, MAE = 1.97 for linear age-bias correction as in Smith et al. (2019), and 2.34 years age-corrected as in Cole et al. (2018).

III. Sample dependance and the effect of training sample size

We additionally compared the performance of the algorithm between different subsamples of the UKB dataset. The subsamples were of comparable sizes and age/sex distributions and showed comparable results (see Table 2). We further tested the effect of varying training sample size by using 20% to 100% of the full dataset that showed only minor improvements in the performance of the algorithm (Figure 1C).

Table 2. *Sample characteristics of different UKB subsamples with MAEs of the mean-ensembled model performance and Pearson's correlation coefficients between the predicted and chronological age.*

Sample	$N_{training}$	$N_{testing}$	Males/Females	Age range	M ± SD	MAE [years]	r_{Age}
UKB 1	6517	724	3503/3738	45–82	64.08±7.54	**2.46**	**0.93**
UKB 2	6745	749	3618/3876	46–82	64.10±7.50	2.47	0.92
UKB 3	6564	729	3406/3887	45–81	64.15±7.53	2.51	0.92
UKB 4	6665	740	3523/3882	44–81	64.05±7.58	2.48	0.92
UKB 5	6666	741	3507/3900	45–81	**64.05±7.52**	2.50	0.92
UKB full	33156	3684	17557/19283	45–82	64.09±7.53	2.35	0.93

IV. Cross-scanner prediction reliability

We investigated the cross-site prediction by training the algorithm on a sample of 6631 T1-weighted MRI scans (using 90% of the sample to keep the size of training sample constant) from scanning site 1 (Cheadle, M_{Age} = 63.45±7.44, age range: 45–81) and tested its performance on a dataset from scanning site 3 (Newcastle, M_{Age} = 64.86±7.45, age range: 47–81), and vice versa.
When predicting from scanning site 1 to 3, we obtained a MAE of 2.4 years (r_{Age} = 0.93), while the reverse resulted in a MAE of 2.48 years (r_{Age} = 0.92).

V. External validation

Furthermore, the models trained on a UKB subsample (UKB 1) were used to predict the BrainAGE in the ADNI sample of healthy subjects, subjects with stable MCI, subjects who later progressed from MCI to AD, and current AD patients. The ADNI test and training samples were matched with regard to sex and age distribution. The results are presented in Figure 1D.

One-way ANOVA showed statistically significant differences between the mean BrainAGE scores of the four groups ($F(3, 691) = 27.073$, $p < .001$). Post hoc comparisons using Tukey HSD test showed significant differences in mean BrainAGE scores of healthy controls and subjects with AD (2.22±0.85 years, $p<.001$), healthy controls and pMCI (2.60±0.90 years, $p<.001$), as well as sMCI and pMCI (2.37±1.12 years, $p<.001$) and sMCI and AD patients (1.99±1.08 years, $p<.001$) at baseline.

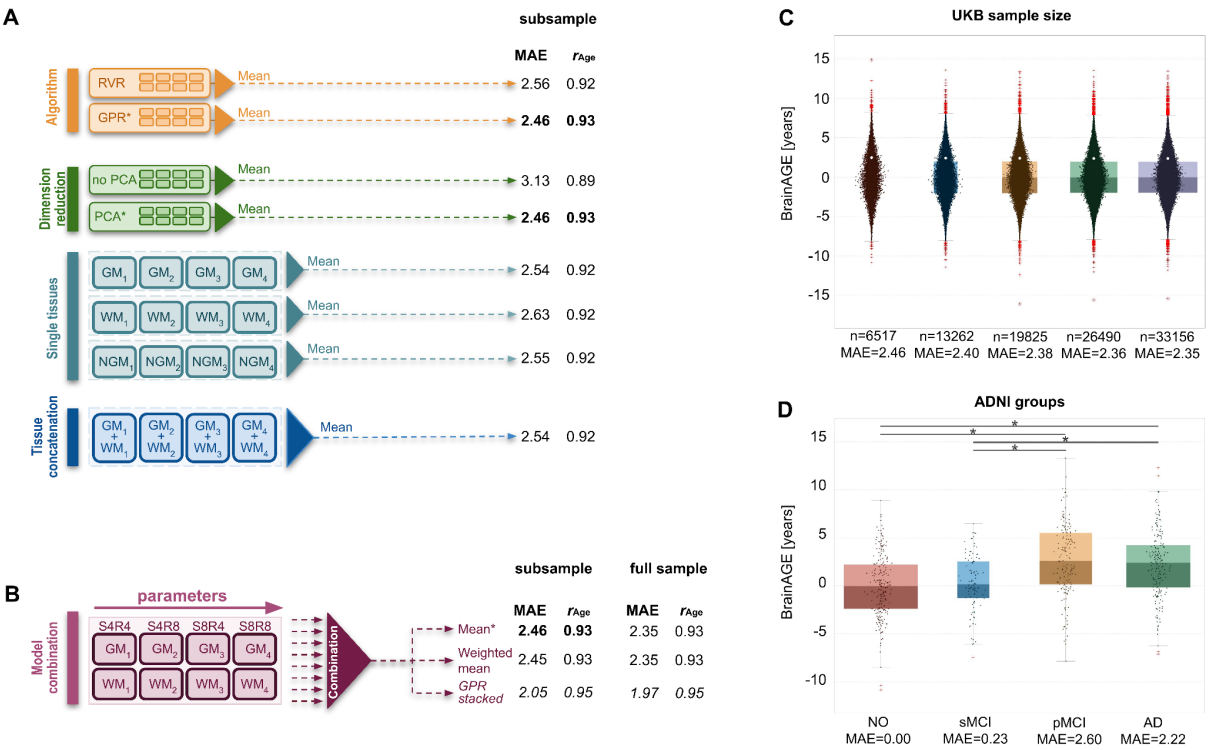

Figure 1

A) *Various single models differing in spatial resolution (R: 4 mm / 8 mm), smoothing kernel (S: 4 mm / 8 mm), and tissue segments (GM, WM, NGM: non-linearly registered modulated GM) were estimated and combined. The results of model estimations combined by averaging are presented as age-bias corrected MAEs and the Pearson's correlation coefficients between predicted- and chronological age for analyses run on a UKB 1 subsample.The comparison of the ML algorithms (RVR/GPR), as well as the effects of dimension reduction and the choice of brain tissues or their concatenation are shown. The results in bold represent the same combined model.* **B)** *Eight single models were ensembled by averaging, weighted averaging, or GPR stacking. The results are presented for the subsample and the full UKB sample.* **C)** *The effect of different training sample sizes.* **D)** *BrainAGE estimations of the baseline images in the ADNI sample (NO - healthy subjects, sMCI - stable MCI, pMCI - progressive MCI, AD - Alzheimer's disease patients).*

VI. Validation on a synthetic dataset of global neocortical thinning

We additionally applied the UKB-trained algorithm to a synthetic dataset of global neocortical thinning from Rusak et al. (2022). The authors simulated neocortical thickness loss from 0 to 0.1 mm in steps of 0.01 mm per time point. The step reflects a yearly rate of cortical loss after the age of 70 (Fox & Schott, 2004; Fjell et al. 2009). They further simulated the hypothetical loss from 0.1 to 1 mm in steps of 0.1 mm which would altogether represent approximately 100 years of global neocortical thinning.

Our aim was to examine how the BrainAGE algorithm performs on simulated data with well-defined neocortical atrophy. We wanted to test whether the BrainAGE algorithm utilises the

global neocortical thinning pattern for age prediction, and see to what extent the BrainAGE scores reflect the assumed yearly changes in neocortical thickness (Figure 2).

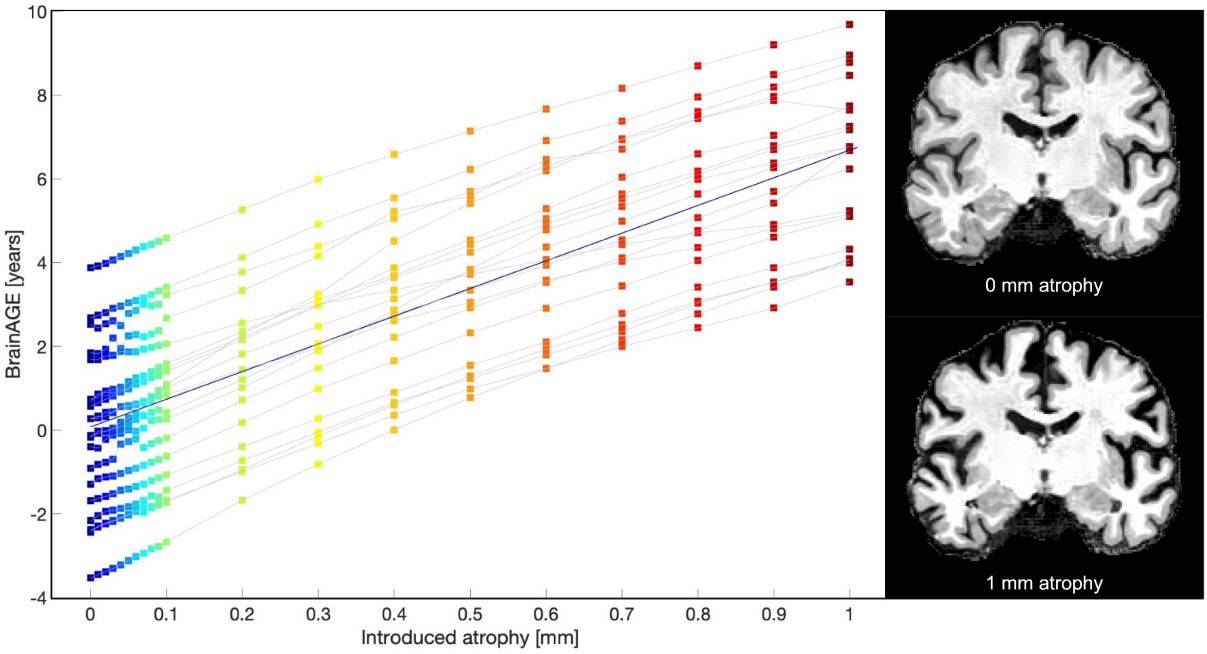

Figure 2. BrainAGE scores of the simulated neocortical thinning dataset with 0 and 1 mm atrophy examples on the right side.

As can be seen from the figure, BrainAGE scores reflect the loss of global cortical thickness, although the values do not represent the expected yearly rate of cortical thinning as only neocortical atrophy was simulated in the dataset.

BrainAGE adjustment for training-sample characteristics

The brain age measures are dependent on the characteristics of the training sample (de Lange et al. 2020; de Lange et al. 2022). Furthermore, with betterment of the prediction accuracy, meaningful deviations between healthy controls and patients decrease (Bashyam et al. 2020). In order to make the results less dependent on training sample characteristics and to improve the differentiation of groups, we performed an adjustment of each group by multiplying the BrainAGE scores with an arbitrary value of 5 (to render the results more age-like) and dividing them by the training MAE. We additionally applied a linear age-bias correction as proposed in Smith et al. (2019) and as in Cole et al. (2018).

To test the effect of the adjustment, we trained the algorithm on the UKB- and the IXI sample, and tested it on the AD patients from the ADNI database. The adjustment with the resulting training MAE produced less varying BrainAGE scores. Furthermore, as can be seen from Figure 3, the adjusted scores are more comparable between the two common age-bias correction approaches.

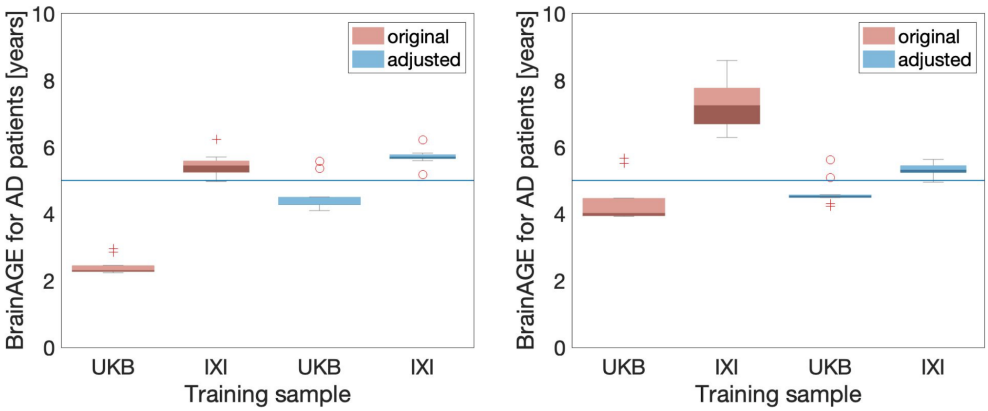

Figure 3. BrainAGE obtained after training on the UKB- or the IXI dataset with age-bias correction (left) as proposed by Smith et al. (2019) and (right) as proposed by Cole et al. (2018) for ADNI AD patients before and after adjustment of the BrainAGE scores by scaling them with 5/(training MAE). This allowed a direct comparison of the age deviation of the AD sample from the healthy ageing population.

Discussion

In our present work, we revised the BrainAGE model that was originally introduced more than a decade ago (Franke et al., 2010). The substitution of RVR with a GPR algorithm proved to be more stable in dealing with a large dataset, as demonstrated by extensive evaluation of the performance of the new framework under different conditions utilising the large neuroimaging sample of the UKB. We explored the contribution of grey and white matter tissue segments and their ensembles to BrainAGE prediction accuracy. By employing the trained model on the ADNI sample, we tested its external validity in dealing with pathological brain tissue changes, whereas the application of the model on a synthetic dataset showed how the algorithm reflects uniform decreases in global neocortical thickness. In addition, we propose an adjusted BrainAGE score approach that makes the values less dependent on specific properties of the training dataset. We discuss the advantages and limitations of our new BrainAGE framework in the following sections.

Dimension reduction & Training sample size

The results showed that one of the most important ways of improving the brain age gap estimation was the reduction of highly individual features that can be seen as “anatomical noise”. In our case it was conducted by smoothing, downsampling, and finally a PCA. As already noted in Franke et al. (2010), not all voxels of structural images are of equal importance to the brain age estimation, therefore, dimensionality reduction is one of the strategies to obtain more accurate results in case of highly dimensional feature space (Baecker et al. 2021; More et al., 2023). It not only reduces computational costs, but also prevents method-specific over-fitting of machine learning models (Franke & Gaser, 2019).

While RVR shows its superiority over GPR on smaller sample sizes (Baecker et al., 2021), training of a GPR model on a large enough sample yielded slightly better results than the RVR approach. However, increasing training sample size over 13,000 subjects resulted in plateauing of BrainAGE accuracies.

Tissue class combination and model ensembling

Based on our analyses, a combination of models trained on affine registered GM and WM segments with varying preprocessing parameterization jointly contribute to the best prediction. Interestingly, the use of non-linearly registered GM data – typically used in VBM studies – did not result in a better performance in comparison to the models that utilised affine registered GM. We speculate that linear registration provides more spatial information of the individual brain anatomy and its atrophy pattern, being more analogous to raw brain anatomy used for deep learning methods, whereas the segmentation (with its denoising and bias-correction steps) helps to remove protocol-based image properties (Ashburner & Friston, 2000).

An important factor that strongly improved the prediction accuracy of BrainAGE estimation was combining various single BrainAGE models. Model ensembling is an effective way of improving the brain age estimation as it serves as a regularisation step to prevent overfitting and thus improves prediction accuracy (van Veen et al., 2015). Even though we ensembled models that were overall quite similar, we believe this is not a strong drawback since even similar models still provide additional variation (van Veen et al., 2015) and our results show benefits without significantly increased computational costs. As can be seen from our results, employing GPR stacking substantially decreased the MAE.

Additional improvements of age prediction may be possible by ensembling more independent brain age models, for example by using different image modalities, morphometric metrics (e.g. thickness, sulcal depth), or other brain-age frameworks (e.g. Cole et al., 2018; Smith et al., 2019). Although improvements by adding modalities were shown before (e.g. Cole, 2020; Niu et al., 2020), we focused only on broadly available T1 image data and a standard VBM paradigm in order to keep our framework parsimonious. Operating within this paradigm renders BrainAGE less of a black box type of measure, potentially enabling explainable artificial intelligence approaches for BrainAGE measures (Boscolo Galazzo et al., 2022). Furthermore, omitting working with other imaging modalities or surface based measures, despite their potentially superior performance that comes at the cost of higher complexity and computational efforts, enables the BrainAGE algorithm to be widely applicable, also for clinical use.

Generalisability and validity

Our results show a comparable performance of the algorithm based on training and testing on different UKB subsamples. A repeated 10 x 10 CV performance demonstrated high consistency between the predicted BrainAGE scores across different folds and repetitions (Table S3 in the supplementary material). Moreover, the cross-scanner prediction within two UKB sample sites shows consistent performance. Further testing of generalisability of the BrainAGE approach was done using the ADNI sample. The BrainAGE model trained on the UKB sample matching age distribution of the ADNI sample showed meaningful differences between the diagnostic groups.

The application of the algorithm to the synthetic global neocortical atrophy dataset proved that the BrainAGE algorithm interprets cortical thinning as ageing, however, the obtained values do not reflect the expected yearly changes. This discrepancy is probably due to the properties of the synthetic dataset which reflects only the neocortical tissue loss, but does not simulate changes in white matter, ventricles, cerebellum, deep grey matter, hippocampus and

amygdala (Rusak et al., 2022), which are structures that also drive brain age estimation (Levakov et al., 2020). In addition, the simulated images were only provided in MNI space which can also affect BrainAGE prediction as a general offset but not the slope. Despite the artificial character of simulated atrophy pattern, the dataset allows for testing specific structural changes (Aubert-Broche et al., 2006, Lerch et al., 2005, Rusak et al. 2022) and their interaction with the brain age algorithms. Simulated datasets can serve as an alternative approach to explainable AI by providing information on the influence of anatomical characteristics on brain age prediction.

Overall, the revised BrainAGE algorithm provides promising and generalisable results. Future studies might benefit from harmonising data over multiple imaging studies (Pomponio et al. 2020), thus allowing the application of brain age prediction to more diverse data.

Adjusted BrainAGE

As the brain age measures are dependent on the characteristics of the training sample (de Lange et al. 2020; de Lange et al. 2022) and the minimisation of the prediction accuracy can result in a decrease in meaningful deviations between healthy controls and patients (Bashyam et al. 2020), we have additionally proposed an adjustment of the BrainAGE scores to render the results less dependent on the training dataset properties and to better differentiate between the scores of group of interest (e.g., AD patients) and healthy subjects. We have shown that adjusting the scores makes results less influenced by training sample-size, renders the estimated values more comparable to other brain age prediction approaches, and reduces differences induced by age-bias correction method. Moreover, it allows a better distinction between BrainAGE scores from healthy subjects and clinical samples, as it normalises the values of healthy control samples. Nevertheless, further studies are warranted to test the usefulness of the proposed idea.

Conclusion

Our BrainAGE algorithm underwent a thorough revision that enabled its application on the very large UK Biobank dataset. Application of the BrainAGE model on subjects with varying neurocognitive impairment levels showcase the model's clinical validity and the BrainAGE scores from a synthetic dataset reflect a meaningful pattern according to known neocortical thickness loss with age. Future research can profit from the use of synthetic data to better understand the inner workings of ML algorithms.

Acknowledgements

This research has been conducted using data from UK Biobank under the application number 41655.

The work was supported by Carl Zeiss Stiftung as a part of the IMPULS project (IMPULS P2019-01-006), the Federal Ministry of Science and Education (BMBF) under the frame of ERA PerMed (Pattern-Cog ERAPERMED2021-127) and the Marie Skłodowska-Curie Innovative Training Network (SmartAge 859890 H2020-MSCA-ITN2019).

The authors gratefully acknowledge the computing time granted by the JARA Vergabegremium and provided on the JARA Partition part of the supercomputer JURECA at Forschungszentrum Jülich.

Authorship contribution statement

Polona Kalc: Writing – original draft. Visualisation. **Robert Dahnke**: Writing – review & editing, Visualisation, Data curation, Software. **Felix Hoffstaedter**: Data curation, Writing – review & editing. **Christian Gaser**: Conceptualization, Methodology, Formal analysis, Software, Writing – review & editing, Visualisation, Supervision, Funding acquisition.

References

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., & Calhoun, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature Communications*, *12*(1), 353. https://doi.org/10.1038/s41467-020-20655-6

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry--the methods. *NeuroImage, 11*, 805–821. doi:10.1006/nimg.2000.0582

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018

Ashburner, J., & Friston, K. J. (2011). Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. *NeuroImage*, *55*(3), 954–967. https://doi.org/10.1016/j.neuroimage.2010.12.049

Aubert-Broche, B., Evans, A. C., & Collins, D. L. (2006). A new improved version of the realistic digital brain phantom. *NeuroImage 32*, 138–145. https://doi.org/10.1016/j.neuroimage.2006.03.052

Baecker, L., Dafflon, J., Costa, P. F., Garcia-Dias, R., Vieira, S., Scarpazza, C., Calhoun, V. D., Sato, J. R., Mechelli, A., & Pinaya, W. H. L. (2021). Brain age prediction: A comparison between machine learning models using region- and voxel-based morphometric data. *Human Brain Mapping*, *42*(8), 2332–2346. https://doi.org/10.1002/hbm.25368

Bashyam, V. M., Erus, G., Doshi, J., Habes, M., Nasrallah, I. M., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., … Davatzikos, C. (2020). MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain, 143*(7), 2312–2324. https://doi.org/10.1093/brain/awaa160

Beheshti, I., Mishra, S., Sone, D., Khanna, P., & Matsuda, H. (2020). T1-weighted MRI-driven Brain Age Estimation in Alzheimer's Disease and Parkinson's Disease. *Aging and Disease*, *11*(3), 618. https://doi.org/10.14336/AD.2019.0617

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 Journal of Biomedical and Health Informatics, 26(4), 1432–1440. 10.1109/JBHI.2021.3083187.

Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906), 525–533. https://doi.org/10.1038/s41586-022-04554-y

Boscolo Galazzo, I., Cruciani, F., Brusini, L., Salih, A., Radeva, P., Storti, S. F., & Menegaz, G. (2022). Explainable Artificial Intelligence for Magnetic Resonance Imaging Aging Brainprints: Grounds and challenges. *IEEE Signal Processing Magazine*, *39*(2), 99–116. https://doi.org/10.1109/MSP.2021.3126573

Chen, T., & Ren, J. (2009). Bagging for Gaussian process regression. *Neurocomputing*, *72*(7–9), 1605–1610. https://doi.org/10.1016/j.neucom.2008.09.002

Cole, J. H. (2020). Multimodality neuroimaging brain-age in UK Biobank: Relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of Aging*, *92*, 34–42. https://doi.org/10.1016/j.neurobiolaging.2020.03.014

Cole, J. H., & Franke, K. (2017). Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends in Neurosciences*, *40*(12), 681–690. https://doi.org/10.1016/j.tins.2017.10.001

Cole, J. H., Leech, R., Sharp, D. J., for the Alzheimer's Disease Neuroimaging Initiative. (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, *77*(4), 571–581. https://doi.org/10.1002/ana.24367

Cole, J. H., Ritchie, S. J., Bastin, M. E., Valdés Hernández, M. C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S. E., Zhang, Q., Wray, N. R., Redmond, P., Marioni, R. E., Starr, J. M., Cox, S. R., Wardlaw, J. M., Sharp, D. J., & Deary, I. J. (2018). Brain age predicts mortality. *Molecular Psychiatry*, *23*(5), 1385–1392. https://doi.org/10.1038/mp.2017.62

Cuadra, M. B., Cammoun, L., Butz, T., Cuisenaire, O., & Thiran, J.-P. (2005). Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. *IEEE Transactions on Medical Imaging*, *24*(12), 1548–1565. https://doi.org/10.1109/TMI.2005.857652

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. 43*(10), 3113-3129. doi: https://doi.org/10.1002/hbm.25837

de Lange, A. G., Anatürk, M., Suri, S., et al. (2020). Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study. *NeuroImage 222*, 117292. https://doi.org/10.1016/j.neuroimage.2020.117292

Eickhoff, C. R., Hoffstaedter, F., Caspers, J., Reetz, K., Mathys, C., Dogan, I., Amunts, K., Schnitzler, A., Eickhoff, S. B. (2021). Advanced brain ageing in Parkinson's disease is related to disease duration and individual impairment. Brain Communications, 3(3), 1–12. https://doi.org/10.1093/braincomms/fcab191

Fjell, A. M., Walhovd, K. B., Fennema-Notestine, C., McEvoy, L. K., Hagler, D. J., Holland, D., Brewer, J. B., Dale, A. M. (2009). One-year brain atrophy evident in healthy aging. *J Neurosci, 29*(48), 15223-31. doi: 10.1523/JNEUROSCI.3252-09.2009.

Fox, N. C., & Schott, J. M. (2004). Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet, 363*, 392–394. doi: 10.1016/S0140-6736(04)15441-X

Franke, K., Clarke, G. D., Dahnke, R., Gaser, C., Kuo, A. H., Li, C., Schwab, M., & Nathanielsz, P. W. (2017). Premature Brain Aging in Baboons Resulting from Moderate Fetal Undernutrition. *Frontiers in Aging Neuroscience*, *9*. https://doi.org/10.3389/fnagi.2017.00092

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. https://doi.org/10.1024/1662-9647/a000074

Franke, K., & Gaser, C. (2019). Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? *Frontiers in Neurology*, *10*, 789. https://doi.org/10.3389/fneur.2019.00789

Franke, K., Gaser, C., Manor, B., & Novak, V. (2013). Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Frontiers in Aging Neuroscience*, *5*. https://doi.org/10.3389/fnagi.2013.00090

Franke, K., Luders, E., May, A., Wilke, M., & Gaser, C. (2012). Brain maturation: Predicting individual BrainAGE in children and adolescents using structural MRI. *NeuroImage*, *63*(3), 1305–1312. https://doi.org/10.1016/j.neuroimage.2012.08.001

Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (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. https://doi.org/10.1016/j.neuroimage.2010.01.005

Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., & Luders, E. (2022). *CAT - A Computational Anatomy Toolbox for the Analysis of Structural MRI Data* [Preprint]. biorXiv. https://doi.org/10.1101/2022.06.11.495736

Gong, W., Beckmann, C. F., Vedaldi, A., Smith, S. M., & Peng, H. (2020). *Optimising a Simple Fully Convolutional Network (SFCN) for accurate brain age prediction in the PAC 2019 challenge* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.11.10.376970

Hahn, T., Fisch, L., Ernsting, J., Winter, N. R., Leenings, R., Sarink, K., Emden, D., Kircher, T., Berger, K., & Dannlowski, U. (2021). From 'loose fitting' to high-performance, uncertainty-aware brain-age modelling. *Brain*, *144*(3), e31–e31. https://doi.org/10.1093/brain/awaa454

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., … ADNI Study. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging, 27(4), 685–691. https://doi.org/10.1002/jmri.21049

Jülich Supercomputing Centre. (2018). JURECA: Data Centric and Booster Modules implementing the Modular Supercomputing Architecture at Jülich Supercomputing Centre Journal of large-scale research facilities, 7, A182. http://dx.doi.org/10.17815/jlsrf-7-182

Levakov, G., Rosenthal, G., Shelef, I., Raviv, T. R., & Avidan, G. (2020). From a deep learning model back to the brain—Identifying regional predictors and their relation to aging. *Hum. Brain Mapp., 41*(12), 3235–3252. doi: 10.1002/hbm.25011

Lombardi, A., Monaco, A., Donvito, G., Amoroso, N., Bellotti, R., & Tangaro, S. (2021). Brain Age Prediction With Morphological Features Using Deep Neural Networks: Results From Predictive Analytic Competition 2019. *Frontiers in Psychiatry*, *11*, 619629. https://doi.org/10.3389/fpsyt.2020.619629

Lerch, J. P. & Evans, A. C. (2005). Cortical thickness analysis examined through power analysis and a population simulation. *NeuroImage, 24*, 163–173. https://doi.org/10.1016/j.neuroimage.2004.07.045

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. https://doi.org/10.1016/j.neuroimage.2016.04.007

Manjon, J., Coupe, P., Marti-Bonmati, L., Collins, D.L., & Robles, M. (2010). Adaptive Non-Local Means Denoising of MR Images With Spatially Varying Noise Levels. *Journal of Magnetic Resonance Imaging, 31*, 192–203. 10.1002/jmri.22003

More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S. B., & Patil, K. R. (2023). Brain-age prediction: A systematic comparison of machine learning workflows. *NeuroImage*, *270*, 119947. https://doi.org/10.1016/j.neuroimage.2023.119947

Nenadić, I., Dietzek, M., Langbein, K., Sauer, H., & Gaser, C. (2017). BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. *Psychiatry Research: Neuroimaging*, *266*, 86–89. https://doi.org/10.1016/j.pscychresns.2017.05.006

Niu, X., Zhang, F., Kounios, J., & Liang, H. (2020). Improved prediction of brain age using multimodal neuroimaging data. *Human Brain Mapping*, *41*(6), 1626–1643. https://doi.org/10.1002/hbm.24899

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., & Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, *68*, 101871. https://doi.org/10.1016/j.media.2020.101871

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Wolf, D. H., … Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, *208*, 116450. https://doi.org/10.1016/j.neuroimage.2019.116450

Rajapakse, J. C., Giedd, J. N., & Rapoport, J. L. (1997). Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Transactions on Medical Imaging*, *16*(2), 176–186. https://doi.org/10.1109/42.563663

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Rogenmoser, L., Kernbach, J., Schlaug, G., & Gaser, C. (2018). Keeping brains young with making music. *Brain Structure and Function*, *223*(1), 297–305. https://doi.org/10.1007/s00429-017-1491-2

Rusak, F., Santa Cruz, R., Lebrat, L., Hlinka, O., Fripp, J., Smith, E., Fookes, C., Bradley, A. P., & Bourgeat, P. (2022). Quantifiable brain atrophy synthesis for benchmarking of cortical thickness estimation methods. *Medical Image Analysis, 82*, 102576. https://doi.org/10.1016/j.media.2022.102576

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. https://doi.org/10.1016/j.neuroimage.2019.06.017

Tohka, J., Zijdenbos, A., & Evans, A. (2004). Fast and robust parameter estimation for statistical partial volume models in brain MRI. *NeuroImage*, *23*(1), 84–97. https://doi.org/10.1016/j.neuroimage.2004.05.007

van Veen, H. J., The Dat, L. N., Segnini, A. (2015). Kaggle Ensembling Guide. https://mlwave.com/kaggle-ensembling-guide/

Varzandian, A., Razo, M. A. S., Sanders, M. R., Atmakuru, A., & Di Fatta, G. (2021). Classification-Biased Apparent Brain Age for the Prediction of Alzheimer's Disease. Frontiers in Neuroscience, 15, 673120. https://doi.org/10.3389/fnins.2021.67312

Supplementary material

Table S1

Mean absolute errors (and Pearson's corr. coefficient between predicted and chronological age) for different age-bias corrections of the BrainAGE estimated on the UKB 1 subsample.

	No age-bias correction	*Linear age-bias correction (Smith et al., 2019)*	*Age-bias correction (Cole, 2020)*
Mean	3.03 (0.86)	2.46 (0.93)	3.48 (0.86)
Weighted mean	3.02 (0.87)	2.45 (0.93)	3.47 (0.87)
GPR stacking	2.26 (0.93)	2.05 (0.95)	2.44 (0.93)

Table S2

Mean absolute errors (and Pearson's corr. coefficient between predicted and chronological age) for different age-bias corrections of the BrainAGE estimated on the full sample.

	No age-bias correction	*Linear age-bias correction (Smith et al., 2019)*	*Age-bias correction (Cole, 2020)*
Mean	2.80 (0.89)	2.35 (0.93)	3.14 (0.89)
Weighted mean	2.80 (0.89)	2.35 (0.93)	3.13 (0.89)
GPR stacking	2.19 (0.93)	1.97 (0.95)	2.34 (0.93)

Table S3

Mean absolute errors and Pearson's correlation coefficient between brain age estimation and chronological age averaged across all CV folds for 10 repetitions (in the repeated 10 x 10-fold CV).

Repetition	*MAE [years]*	r_{Age}
1	*2.45*	*0.925*
2	*2.47*	*0.925*
3	*2.46*	*0.925*
4	*2.47*	*0.925*
5	*2.45*	*0.926*
6	*2.45*	*0.926*
7	*2.45*	*0.925*
8	*2.44*	*0.926*
9	*2.46*	*0.925*
10	*2.44*	*0.926*