NeuroImage 285 (2024) 120469

Contents lists available at ScienceDirect

NeuroImage

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

BASE: Brain Age Standardized Evaluation ☆

Lara Dular [a], Žiga Špiclin [a,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] University of Ljubljana, Faculty of Electrical Engineering, Tržaška cesta 25, Ljubljana, 1000, Slovenia

ARTICLE INFO

Keywords:
Brain age
Evaluation
Deep regression
Accuracy
Robustness
Reproducibility
Consistency
UK biobank

ABSTRACT

Brain age, most commonly inferred from T1-weighted magnetic resonance images (T1w MRI), is a robust biomarker of brain health and related diseases. Superior accuracy in brain age prediction, often falling within a 2–3 year range, is achieved predominantly through deep neural networks. However, comparing study results is difficult due to differences in datasets, evaluation methodologies and metrics. Addressing this, we introduce Brain Age Standardized Evaluation (BASE), which includes (i) a standardized T1w MRI dataset including multi-site, new unseen site, test-retest and longitudinal data, and an associated (ii) evaluation protocol, including repeated model training and upon based comprehensive set of performance metrics measuring accuracy, robustness, reproducibility and consistency aspects of brain age predictions, and (iii) statistical evaluation framework based on linear mixed-effects models for rigorous performance assessment and cross-comparison. To showcase BASE, we comprehensively evaluate four deep learning based brain age models, appraising their performance in scenarios that utilize multi-site, test-retest, unseen site, and longitudinal T1w brain MRI datasets. Ensuring full reproducibility and application in future studies, we have made all associated data information and code publicly accessible at https://github.com/AralRalud/BASE.git.

1. Introduction

Brain age is an estimate of biological age derived from brain magnetic resonance images (MRIs), and it has emerged as a significant biomarker of neurological health and aging. Assessing brain age involves training a machine learning model for age prediction using input T1-weighted (T1w) MRIs of a healthy population, followed by the application of the model outside the training dataset to detect potential brain age discrepancies in diverse health conditions. For instance, increased brain age with respect to healthy controls has been demonstrated in patients with neurological diseases such as Alzheimer's dementia (Franke and Gaser, 2012), multiple sclerosis (Høgestøl et al., 2019; Cole et al., 2020), schizophrenia (Schnack et al., 2016; Koutsouleris et al., 2014), and other diseases like type 2 diabetes (Franke et al., 2013), human immunodeficiency virus (HIV) (Petersen et al., 2021; Cole et al., 2017c), and in obese (Ronan et al., 2016) and vitamin D deficient subjects (Terock et al., 2022).

The use of deep learning (DL) models for brain age prediction has seen a surge in recent years (Baecker et al., 2021b; Tanveer et al., 2023). However, differences in evaluation protocols, such as the use of varying performance metrics, different validation datasets, age spans, subject counts, T1w preprocessing pipelines, and post-processing age-bias corrections, make comparisons across studies challenging, if not impossible. Although the evaluation of models on new site data is somewhat common, their evaluation on longitudinal datasets to assess the ability to capture the linear trend associated with aging, is rather rare. Even in studies that performed such evaluations (Dartora et al., 2022; Dunås et al., 2021; Beheshti et al., 2021), the consistency of predictions was either assessed visually or based on cross-sectional metrics, which seems inadequate. Furthermore, the reproducibility of predictions across models trained with different weight initializations (Jonsson et al., 2019; Levakov et al., 2020) or those using test-retest settings (Franke and Gaser, 2012; Cole et al., 2017b; Feng et al., 2020) has not been systematically evaluated.

To bridge these gaps, we propose the Brain Age Standardized Evaluation (BASE), which aims to establish a standardized approach to evaluate brain age prediction models, integrating best practices and overcoming the limitations of existing methodologies.

☆ This study was supported by the Slovenian Research Agency (Core Research Grant No. P2-0232 and Research Grants Nos. J2-2500 and J2-3059). The APC was funded by the Slovenian Research Agency.

* Corresponding author.
E-mail address: ziga.spiclin@fe.uni-lj.si (Ž. Špiclin).

[1] 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: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.neuroimage.2023.120469
Received 27 July 2023; Received in revised form 31 October 2023; Accepted 20 November 2023
Available online 6 December 2023

This paper is organized as follows: a review of related work is provided in Section 2; Section 3 describes the BASE datasets, performance metrics, and evaluation protocols, and a statistical framework used for assessing brain age models; the models and their evaluation using BASE are detailed in Sections 4 and 5, respectively; finally, the discussion and conclusion are presented in Sections 6 and 7, respectively.

2. Related work and contribution

Recent research efforts in brain age prediction have focused on introducing novel DL architectures (He et al., 2022b,a; Bellantuono et al., 2021), diversifying training strategies, including cascade learning (Cheng et al., 2021) and model ensembling over modalities (Kuo et al., 2021; Peng et al., 2021; Dunås et al., 2021; Jonsson et al., 2019), modifying the input T1w image into a two-channel representation encoding contrast and morphometry information (He et al., 2021), simplifying preprocessing by utilizing only image registration to common space (Dartora et al., 2022), and optimizing sampling strategies, to achieve an evenly sampled training set over the entire age span (Feng et al., 2020). A general deficiency of these research studies is the lack of a common, standardized evaluation approach.

Present methodologies for evaluating brain age models predominantly concentrate on contrasting the performances of traditional machine learning models (Beheshti et al., 2022; Baecker et al., 2021a; Han et al., 2022; Xiong et al., 2023). In these studies, models are typically trained and tested on the same collection of MRIs. Such evaluations may fall short in fully capturing various confounding elements such as subject and scanner variability, thus sidelining several crucial aspects of model performance. Although the recent comprehensive research by More et al. (2023) delves into these aspects, it primarily focuses on traditional machine learning models. It thereby overlooks certain aspects intrinsic to deep learning models, such as the reproducibility of predictions of multiple models trained with different weight initializations and the effect of potential alterations in preprocessing between training and test datasets.

The *accuracy* of brain age models is conventionally assessed through the Mean Absolute Error (MAE) computed across all test subjects, signifying the discrepancy between biological and predicted age. However, MAE can present a misleading picture, particularly when the test data comprises age ranges that are overrepresented in the training data, leading to more precise predictions (for instance, when there is a high proportion of young subjects as in the OpenBHB dataset (Dufumier et al., 2021)). As such, the MAE is not sensitive to the possible increase (or decrease) of absolute errors in specific age subintervals. Some studies attempt to circumvent this issue by reporting the MAE by age interval (He et al., 2022b; Levakov et al., 2020; Amoroso et al., 2019). There is a clear need for a *robustness* metric to differentiate between close-fitting (Cheng et al., 2021) models, which demonstrate consistent precision across all ages, and loose-fitting (He et al., 2021) models, which exhibit variable accuracy, especially in underrepresented age intervals throughout the entire age span.

Methodological studies reporting improvements in brain age prediction accuracy on healthy subjects often lack rigorous statistical evaluation. Conversely, studies on diseased populations typically involve statistical evaluation, employing t-test and/or ANOVA with post hoc pairwise comparisons (Franke and Gaser, 2012). Noteworthy practices include the use of Linear Mixed-effects Models (LMEM) on subjects with Alzheimer's disease, mild cognitive impairment, schizophrenia or depression (Bashyam et al., 2020), and multiple sclerosis (Høgestøl et al., 2019; Cole et al., 2020), using the brain age gap as an independent variable. Such a rigorous statistical framework and its parametrization, is yet to be established for evaluation of brain age on healthy subject datasets.

Validation of brain age prediction models for clinical applications should involve assessing their performance on new (unseen) site T1w subject scans, not used during model training (Feng et al., 2020; Jonsson et al., 2019; Dufumier et al., 2022; Franke and Gaser, 2012; He et al., 2021, 2022b,a; Bellantuono et al., 2021; Han et al., 2022; Dartora et al., 2022; Cai et al., 2023; Bashyam et al., 2020). When models are applied to an unseen dataset, a deterioration in performance metrics is generally observed (Feng et al., 2020; Dufumier et al., 2022; Jonsson et al., 2019; Han et al., 2022; Dartora et al., 2022; Cai et al., 2023; Bashyam et al., 2020), but which is often compensated for through a linear bias linear correction. However, a recent study advises against such age-bias correction, since bias corrected metrics can indicate high accuracy, even for models showing poor initial performance (de Lange et al., 2022; Butler et al., 2021). Nevertheless, when the offset appears to be systematic across the entire age span (Franke and Gaser, 2012), applying an offset adjustment may be appropriate.

The *consistency* of age predictions is vital for longitudinal intra-subject evaluations, especially when tracking disease progression or deviations from the normative aging trajectory. While there has been significant progress in providing extensive public datasets and benchmarking platforms, which incorporate multi-site train and test datasets, as well as new site data (for instance, the OpenBHB (Dufumier et al., 2022)), research on longitudinal datasets involving healthy subjects remains underrepresented. Current studies usually resort to visual methods to evaluate longitudinal consistency through charting longitudinal predictions on linear graphs (Dunås et al., 2021; Dartora et al., 2022). Quantitative longitudinal performance evaluation metrics were used in the study by Dunås et al. (2021), where linear lines between time points were computed to analyze the longitudinal predicted trajectories. While the analysis of slope and intercept allows monitoring the rate of change over time, it does not capture the information on the magnitude of the error of the predicted difference, that would be analogous to MAE. This observation underscores the necessity for specialized metrics designed to evaluate the consistency of brain age predictions on longitudinal data.

Finally, the *reproducibility* of any biomarker holds vital importance for practical application and can be assessed using test-retest data. However, brain age studies have thus far used either (i) a limited number of test-retest subjects with a large number of scans per subject (Feng et al., 2020) or (ii) a large number of test-retest subjects, each with few scans (Cole et al., 2017b; Franke and Gaser, 2012). The best observed practice for assessing test-retest agreement is to report the intraclass correlation coefficient (ICC). Another aspect is the reproducibility of brain age predictions across DL model realizations, considering the initial random weight selection, where ICC can also be utilized. However, such evaluations have rarely been performed in the studies involving DL models (Jonsson et al., 2019; Levakov et al., 2020).

The contribution of this paper is BASE, which comprises (i) a standardized T1w MRI dataset including multi-site, new unseen site, test-retest, and longitudinal datasets, along with (ii) an evaluation protocol. The evaluation protocol includes a comprehensive set of established and novel performance metrics to measure the accuracy, robustness, reproducibility, and consistency aspects of brain age predictions, complemented by a statistical evaluation framework based on LMEMs. This protocol is crafted for compatibility not just with our proposed T1w MRI dataset, but can also be adapted for use with alternative datasets relevant to brain age prediction. We demonstrate the use of BASE in a comprehensive evaluation of four DL brain age models, with reproducible results using our public implementation at https://github.com/AralRalud/BASE.git.

3. BASE protocol

The BASE protocol, depicted in Fig. 1, outlines tasks in the model training and tuning, and model evaluation phases. The former involves model training, hyperparameter tuning, repeated model training with different weight initializations and prediction ensembling.

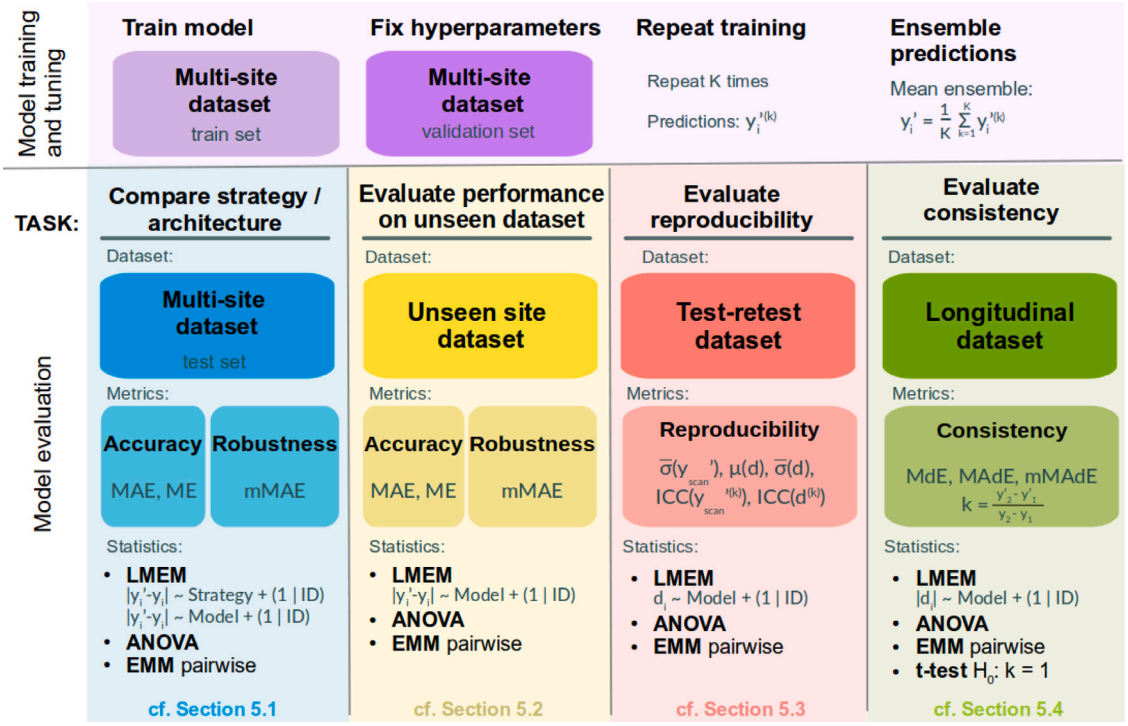

Fig. 1. The BASE protocol involves model training, tuning (*top*), and model evaluation phase (*bottom*), each encompassing specific tasks.

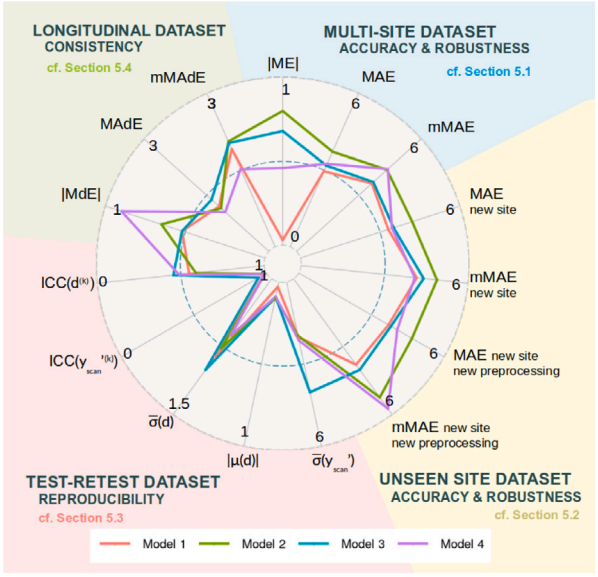

Fig. 2. The principal results of BASE are visualized in the form of a radar plot. Values closer to the plot's center indicate better performance, therefore a tighter envelope indicates a better overall performance for a particular model.

The model evaluation phase involves four tasks: (1) comparison of the performance of DL models and/or the comparative evaluation of the impact of model training strategies, (2) performance evaluation on seen/unseen dataset, (3) reproducibility and (4) consistency evaluation on respective test-retest and longitudinal datasets. The principal results of BASE, sourced from Sections 5.1–5.4, are depicted in the form of a radar plot in Fig. 2.

The building blocks of BASE comprise the data, performance metrics, and statistical analysis framework, which are detailed in the following subsections.

3.1. Datasets

In developing BASE, we established four distinct datasets. The primary dataset encompasses multi-site T1w MRIs, allocated for purposes of training, validation, and testing. The remaining three datasets are dedicated exclusively to testing, each serving a specific function: one for new unseen site T1w MRIs, another for test-retest T1w MRIs, and the last for longitudinal T1w MRIs. Across all datasets, the included subjects are healthy adults, ranging in age from 18 to 95 years old.

The **multi-site dataset** (cf. Table 1) comprises seven publicly available datasets and included a total of 4428 T1w MRIs of healthy subjects. Many of these datasets sourced their images from several hospitals or sites, employing a variety of MRI scanners, such as GE, Siemens, and Philips, with 1.5T and 3T field strengths. OASIS 2 and CamCAN datasets were the only datasets in which scans were acquired on a single scanner. The incorporation of these datasets from multiple sources, sites, and vendors inherently leads to variations in the acquisition pipelines.

All MRIs underwent a visual quality check. Images that did not pass the visual quality check (e.g. due to motion artifacts) were excluded ($N_{excl} = 408$), while subjects under the age of 18 or with non-disclosed ages were discarded ($N_{disc} = 481$). For subjects with multiple T1w scans, we retained the chronologically first non-discarded image. Ultimately, 2504 T1w MRIs were accepted and split into training ($N = 2012$), validation ($N = 245$) and test ($N = 247$) datasets. The distribution of subjects' ages per dataset, as well as within the train/validation/test subsets, is provided in the Supplementary Materials (Appendix A.1). For reproducibility purposes, we have made the subject IDs for each split available in the online project repository.[2]

The **unseen site** and **longitudinal dataset** were sourced from a subset of UK Biobank (UKB) dataset (Miller et al., 2016). We identified 1493 subjects who met the inclusion criteria, which included having two MR scans and no long-standing illnesses. In addition, subjects were required to self-report an overall health rating of *excellent* or *good* at

[2] https://github.com/AralRalud/BASE

Table 1
Dataset information including age statistics, such as span, mean age (μ_{age}), and associated standard deviation ($\mathbf{sd}_{age}$) in *years*, is provided per dataset for the included T1w subject scans in train, test, and validation datasets (*top*), and as well as the new unseen site, test-retest and longitudinal datasets (*bottom*).

Aim: Train, Validation, Test (Multi-site dataset; for evaluating accuracy and robustness)					
Dataset		$\mathbf{N}_{scans}$	**M%/F%**	**Age span**	$\mu_{age} \pm \mathbf{sd}_{age}$
ABIDE I[a]		161	88/12	18.0 – 48.0	25.7 ± 6.4
ADNI[b,i]		248	51/49	60.0 – 90.0	76.2 ± 5.1
CamCAN (Shafto et al., 2014; Taylor et al., 2017)[c]		624	49/51	18.0 – 88.0	54.2 ± 18.4
CC-359 (Souza et al., 2018)[d]		349	49/51	29.0 – 80.0	53.5 ± 7.8
FCON 1000[e]		572	34/66	18.0 – 85.0	45.3 ± 18.9
IXI[f]		472	47/53	20.1 – 86.2	49.0 ± 16.2
OASIS-2 (Marcus et al., 2010)[g]		78	28/72	60.0 – 95.0	75.6 ± 8.4
Total		2504	48/52	18.0 – 95.0	52.1 ± 19.1
Aim: Unseen site dataset (for evaluating the generalization of accuracy and robustness)					
Dataset	$\mathbf{N}_{subj}$	$\mathbf{N}_{scans}$		**Age span**	$\mu_{age} \pm \mathbf{sd}_{age}$
UK Biobank (Miller et al., 2016)	1493	1493	45/55	48.5 – 80.4	63.2 ± 7.2
Aim: Test-retest dataset (for evaluating reproducibility)					
Dataset	$\mathbf{N}_{subj}$	$\mathbf{N}_{scans}$		Age span	$\mu_{age} \pm \mathbf{sd}_{age}$
OASIS-1 (Marcus et al., 2007)[h]	316	632	38/62	18.0 – 94.0	45.1 ± 23.9
Aim: Longitudinal dataset (for evaluating consistency)					
Dataset	$\mathbf{N}_{subj}$	$\mathbf{N}_{scans}$		**Age span**	$\mu_{age} \pm \mathbf{sd}_{age}$
UK Biobank (Miller et al., 2016)	1493	2986	45/55	48.5 – 82.7	64.3 ± 7.2

Dataset information and download at:

[a] ABIDE I: http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html.

[b] ADNI: http://adni.loni.usc.edu/.

[c] CamCAN: https://camcan-archive.mrc-cbu.cam.ac.uk/dataaccess/.

[d] CC-359: https://sites.google.com/view/calgary-campinas-dataset/download.

[e] FCON 1000: http://fcon_1000.projects.nitrc.org/indi/enhanced/neurodata.html.

[f] IXI: https://brain-development.org/ixi-dataset/.

[g] OASIS-2: https://www.oasis-brains.org/.

[h] OASIS-1: https://www.oasis-brains.org/.

[i] 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 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).

both scans. The unseen site dataset comprises 1493 baseline scans, while the longitudinal dataset includes 2986 T1w MRI scans from both baseline and follow-up sessions. The average time between scans was 2.25 ± 0.12 years.

Finally, for the **test-retest datasets**, we used the OASIS-1 dataset (Marcus et al., 2007), which comprises 316 healthy adults, each with two T1w MRIs acquired within a couple of hours (i.e., test-retest scans).

3.2. Performance metrics

The established metric for assessing *accuracy* in model predictions is the mean absolute error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i' - y_i\right|,$$

where y_i denotes true age and $y_i\}$ predicted age of ith subject. Additionally, we report the mean error (ME):

$$ME = \frac{1}{N}\sum_{i=1}^{N}\left(y_i' - y_i\right),$$

to detect instances of age under- or over-estimation, with the assumption that ME is normally distributed around zero mean.

Model is considered *robust* if the MAE within any age subinterval is consistent with the overall MAE. To assess this, we propose the maximal MAE (mMAE), which is calculated by taking the maximum MAE over age intervals $[c_0, c_1)$, $[c_1, c_2)$, …, $[c_{M-1}, c_M]$, as:

$$mMAE = \max_j \frac{1}{N_j}\sum_{i=1}^{N_j}\mathbb{1}_{\{y_i \in [c_j, c_{j+1})\}}\left|y_i' - y_i\right|,$$

where N_j is the number of samples from the interval $[c_j, c_{j+1})$ and $\sum_j N_j = N$. To ensure a balanced number of subjects in each interval on the test set, we utilized intervals $[18, 25], (25, 35], \ldots, (75, 85], (85, 100]$.

A model exhibits high *reproducibility* if the prediction error is similar on test-retest scans and/or across models of the same architecture trained with different weight initializations. We compute the average standard deviation (SD) of the scan predictions:

$$\begin{aligned}\overline{\sigma}(y'_{scan}) &= \frac{1}{tN}\sum_{i=1}^{N}\sum_{j=1}^{t}\sigma(y'_{ij}) \\ &= \frac{1}{tN}\sum_{i=1}^{N}\sum_{j=1}^{t}\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(y_{ij}'^{(k)} - \bar{y}'_{ij})^2}, \\ \bar{y}'_{ij} &= \frac{1}{K}(y_{ij}'^{(1)} + \cdots + y_{ij}'^{(K)})\end{aligned}$$

where N denotes the number of subjects, t the number of scans per subject, and K the number of repeated model trainings, each with different weight initialization. Further, $y_{ij}'^{(k)}$ denotes the prediction of kth model for jth scan of subject i and $\bar{y}'_{ij}$ its mean prediction over all K trained models.

We further computed the mean difference $\mu(d)$ of per subject test-retest predictions as

$$\mu(d) = \frac{1}{N}\sum_{i=1}^{N}\bar{d}'_i$$

and the corresponding average SD of differences $\overline{\sigma}(d)$ as

$$\overline{\sigma}(d) = \frac{1}{N}\sum_{i=1}^{N}\sigma(d_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(d_i^{(k)} - \bar{d}_i')^2},$$

where

$$\bar{d}_i' = \frac{1}{K}(d_i'^{(1)} + \cdots + d_i'^{(K)})$$

denotes the mean subject difference of K trained models, and $d_i'^{(k)} = y_{i2}'^{(k)} - y_{i1}'^{(k)}$ represents the difference in prediction between ith subject's first and second scan for kth model.

The degree of agreement between the predictions of models trained with different weight initializations is further quantified using intra-class correlation coefficient (ICC) (Finn, 1970).

Differences in brain age predictions on successive scans of healthy subjects should be *consistent* with the time elapsed between scan acquisitions. To evaluate this, we compute the age difference between the baseline (B) and follow-up (F) scans for ith subject, $d_i = y_i^F - y_i^B$, and corresponding difference of model predictions, $d_i' = y_i^{F'} - y_i^{B'}$. Subsequently, we define the Mean Difference Error (MdE), Mean Absolute Difference Error (MAdE) and Maximal Mean Absolute Difference Error (mMAdE) as:

$$MdE = \frac{1}{N}\sum_{i=1}^{N}(d_i - d_i'),$$

$$MAdE = \frac{1}{N}\sum_{i=1}^{N}\left|d_i - d_i'\right|,$$

$$mMAdE = \max_j \frac{1}{N_j}\sum_{i=1}^{N_j}\mathbb{1}_{\{y_i\in[c_j,c_{j+1})\}}\left|d_i - d_i'\right|,$$

using the same age interval boundaries $c_1, c_2, \ldots, c_M$ as defined for the *robustness* assessment.

In case the longitudinal and test-retest data contain $t > 2$ images per subject, the formulation of $\mu(d)$, $\overline{\sigma}(d)$, MdE, $MAdE$ and $mMAdE$ can be generalized by averaging across all $\binom{t}{2}$ pairs per subject.

3.3. Statistical analysis

Linear mixed-effects models (LMEMs) were employed to characterize the relationship between the error and the absolute error (AE) as dependent variables, with the model architecture serving as a fixed effect and the subject ID as a random effect. This configuration ensures that all responses from a specific subject are adjusted by a unique additive value corresponding to that subject. By treating the subject ID as a random effect, we effectively accommodated the dependent nature of the data, which stems from generating multiple brain age predictions for the same individual.

For all models, we report the estimated regression coefficients along with their 95% confidence intervals (CIs). To explain the variability in the response variable due to the fixed effect, we performed an Analysis of Variance (ANOVA) on the fitted model and pairwise comparisons between the levels of the fixed factor using the Estimated Marginal Means (EMM) method, with a Tukey adjustment for multiple comparisons.

The LMEM analysis was conducted in R version 4.0.4, using 'lme4' package version 1.1.26. For computing p-values of ANOVA tests, we used package 'lmerTest' version 3.1.3. Finally, pairwise analysis was conducted using package 'emmeans' version 1.5.4.

To statistically evaluate longitudinal consistency, testing whether the average slope between age estimates on baseline and follow-up T1w scans differs from 1 (null hypothesis), we ran the t-test.

In all statistical tests, the significance threshold was set at $\alpha = 0.05$, unless noted otherwise.

4. Brain age prediction

4.1. T1-weighted image preprocessing

Each input T1w image was converted to the Nifti format. The raw T1w image underwent adaptive non-local means denoising[3] (Manjón et al., 2010). Next, we performed a 12 degree-of-freedom affine registration using NiftyReg[4] (Modat et al., 2014) to map the denoised T1w image into the 7th generation Montreal Neurological Institute (MNI) atlas space (version 2009c) (Fonov et al., 2009). To improve registration accuracy, intensity inhomogeneity correction (w/o mask) was applied to the denoised image using the N4 algorithm[5] (Tustison et al., 2010), prior to running the registration. The intensity-inhomogeneity-corrected, denoised T1w image was used during registration only. With the obtained affine mapping, the denoised T1w image was resampled to the MNI space using sinc interpolation, such that all preprocessed T1w images had a size of 193 × 229 × 193 and isotropic 1 mm spacing.

Finally, a two-step grayscale correction was applied: (1) intensity windowing, which involves the computation of the lower and upper thresholds based on the grayscale histogram, smoothed with a Gaussian filter. The lower threshold is set based on the histogram's lowest intensity mode location plus twice the value of the mode's full width at half maximum (FWHM). Note that the particular mode corresponds to the grayscale values of the background and non-tissue regions of the T1w MRI image. To compute the upper threshold, the grayscale values beyond the 99th percentile are first set to the value of the lower threshold. Inflection points in the intensity distribution from the 50th to the 95th percentiles are then identified by computing the second derivative. The upper threshold is defined as the value of the percentile at a selected inflection point, plus three times the Median Absolute Deviation of the pixel intensities that are above the lower threshold. The second step, (2) involves intensity inhomogeneity correction, utilizing the N4 algorithm with the MNI152 atlas mask dilated by 3 voxels. In all the resulting preprocessed T1w MRI images we removed the non-informative empty space around the head by cropping to a size of 157 × 189 × 170.

4.1.1. UK biobank T1w preprocessing

We utilized the UKB dataset, employing both raw T1w MRI images and preprocessed images obtained through the protocol outlined in Smith et al. (2022). We generated two versions of preprocessed T1w MRI images: (1) The raw T1w defaced images in subject image space were preprocessed as described in previous section, and (2) the UKB preprocessed T1w images were spatially mapped from the 6th to 7th generation MNI atlas (version 2009) (Fonov et al., 2009), using 3rd order interpolation and a linear transformation matrix between the two atlases pre-computed by FSL's FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002) to ensure that the preprocessed images were registered to the same atlas as used in our preprocessing.

4.2. Prediction models

To showcase BASE, we reimplemented four CNN-based brain age models based on the descriptions in the literature. The architectures of the four models are depicted in Fig. 3.

Model 1 (Cole et al., 2017b) was among the first 3D regression CNNs applied for brain age prediction, and was trained and tested on the preprocessed T1w MRIs. **Model 2** (Huang et al., 2017) is a multi-channel 2D regression CNN, trained and tested on 15 equidistantly

[3] Adaptive non-local means denoising Version 2.0: https://github.com/djkwon/naonlm3d

[4] NiftyReg Software http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg

[5] N4 bias field correction: https://manpages.debian.org/testing/ants/N4BiasFieldCorrection.1.en.html

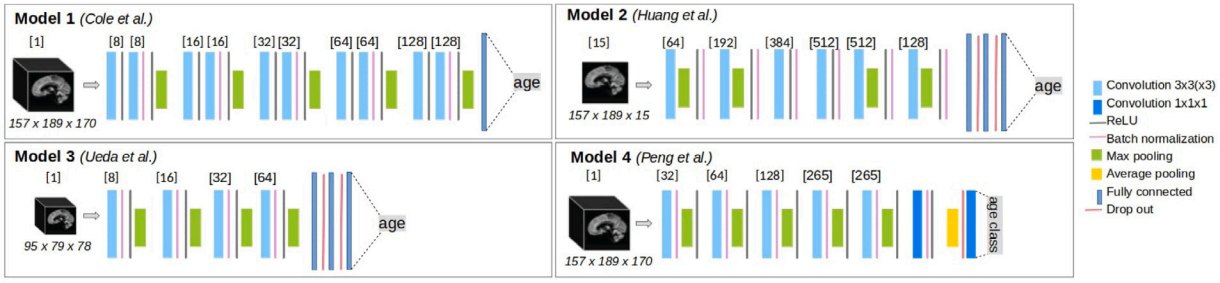

Fig. 3. Architecture of the four reimplemented CNN models for the task of brain age prediction.

sampled axial slices of the preprocessed T1w as input channels. **Model 3** (Ueda et al., 2019) is similar to Model 1, but applied on downsampled 3D T1w. The use of multi-channel 2D or downsampled 3D models may reduce computational complexity with little impact on prediction performance, as motivated by a recent review paper on DL brain age regression (Tanveer et al., 2023); a hypothesis that we aim to verify.

Finally, **Model 4** (Peng et al., 2021) is a fully convolutional classification model, outputting probability over non-overlapping 2-year age intervals, that reported one of the best results for brain age prediction among DL models. It was trained and tested on the preprocessed T1w images, using a weighted sum over the class probabilities to predict the age.

All models were implemented in PyTorch 1.4.0 for Python 3.6.8. The details on model training and hyperparameter tuning are presented in the Supplementary materials A.1.

4.3. Offset correction

Predicting age on a dataset involving domain shift (i.e., unseen scanner and/or T1w preprocessing) usually incurs a drop in accuracy, observed as a systematic offset versus the true age. We applied the offset correction for the value of ME, calculated as:

$$y_i^{corr} = y_i' - ME = y_i' - \frac{1}{N}\sum_{j=1}^{N}(y_j' - y_j).$$

The offset correction was applied to all four models, computed on a per-model basis.

Several recent studies have cautioned against the use of linear age-bias correction (de Lange et al., 2022; Butler et al., 2021). These methods involve regressing the age out of the brain age gap and forcing an alignment between predicted and true age, even for poorly fitting models. In the worst-case scenario, a poorly fitting model would predict the median age for all subjects. Unlike fitting a linear regression line, offset correction does not force this alignment and does not correct the model's inability to capture a linear trend, nor does it reduce dispersion of predictions.

5. Experiments and results

Our experiments showcase an objective, quantitative, and comparative evaluation of the four DL-based brain age models using BASE in four tasks, each with a corresponding set of data, performance metrics, and statistical analyses, as outlined in the following subsections.

5.1. Impact of model architecture

The performance of four DL model architectures, described in Section 4.2, were evaluated. We trained a total of 20 models on the multi-site test set, i.e., $K = 5$ random weight initializations for each of the four models.

The final predictions were obtained by averaging the $K = 5$ predictions across models with different weight initialization. The so-called mean ensembling strategy has been shown to generally improve models' accuracy (Jonsson et al., 2019; Levakov et al., 2020; Peng et al., 2021; Couvy-Duchesne et al., 2020).

	Model 1	Model 2	Model 3	Model 4
Model 1		0.76	0.22	0.29
Model 2	-0.76		-0.54	-0.47
Model 3	-0.22	0.54		0.07
Model 4	-0.29	0.47	-0.07	

Fig. 4. Pairwise differences of EMMs for the LMEM model. Statistical significance is marked in red for $p < 0.001$, orange for $0.001 < p < 0.01$ and yellow for $0.01 < p < 0.05$.

We evaluated the accuracy and robustness of age predictions for the models trained on the multi-site dataset, obtained by the mean ensembling strategy on the multi-site test dataset. We fit a LMEM with AE as the dependent variable, subject ID as a random effect and model architecture as a fixed effect.

Results in Table 2 show that the best accuracy was achieved by the mean ensemble for Model 1, with a MAE of 2.96 years and ME close to zero. Furthermore, the performance of Model 1 versus the other models resulted in comparatively small SDs of ME and MAE. According to the MAE values, as well as their SDs, Models 1, 3, and 4 performed better than Model 2, due to the former inputting 3D T1w MRI, while the latter the inputted subsampled 2D axial slices. The most robust model, according to mMAE, was the Model 1. Among the models inputting 3D T1w MRIs, Model 4 performed the worst in terms of accuracy and robustness. Furthermore, we observed that R^2 and r have little or no discriminating power to differentiate model performances. An overview of the models' performances on the multi-site dataset, including the MAE, mMAE, and the absolute value of ME, is visually presented in the top-right blue area of the radar plot in Fig. 2.

In evaluating the significance of the observed differences, the LMEM analysis and ANOVA test ($F(3, 738) = 7.709$, $p < 0.001$) showed that model architecture had a significant effect on the AE. The exact regression coefficients, their 95% CI, and ANOVA F-values are reported in Supplementary Table 8. The results of LMEM post-hoc pairwise analysis are presented in Fig. 4. The AEs of Model 2 were statistically significantly different from those of Models 1, 3, and 4. The EMMs did not significantly differ for the other model pairs.

5.2. Performance on unseen site dataset

The four models employing the mean ensembling strategy were applied to the unseen site dataset, which utilized two distinct T1w preprocessing procedures: one identical to the preprocessing of the training

Table 2
Evaluation of brain age prediction for four DL models on the multi-site test set. Best metric results with respect to model architecture (in *rows*) are marked in **bold**. All numbers are in years.

	Accuracy		Robustness	RMSE	R^2	r
	ME (SD)	MAE (SD)	mMAE			
Model 1	**−0.03 ± 3.86**	**2.96 ± 2.47**	**3.63**	**3.85**	**0.96**	**0.98**
Model 2	−0.80 ± 4.81	3.72 ± 3.14	4.36	4.87	0.94	0.97
Model 3	−0.68 ± 3.90	3.18 ± 2.35	3.70	3.95	**0.96**	**0.98**
Model 4	−0.46 ± 4.21	3.25 ± 2.70	4.40	4.22	0.95	**0.98**

Table 3
Accuracy and robustness metrics for the previously unseen UKB dataset. The best metric result with respect to the model architecture (in *rows*) are marked in **bold**. All numbers are in years.

	Same preprocessing			**New preprocessing**		
	Accuracy		**Robustness**	**Accuracy**		**Robustness**
	ME (SD)	MAE (SD)	mMAE	ME	MAE	mMAE
Model 1	−2.10 ± 4.21	3.73 ± 2.88	5.12	−3.33 ± 4.64	4.65 ± 3.31	**5.19**
Model 2	**−1.58 ± 5.53**	4.32 ± 3.78	5.88	−9.58 ± 5.84	9.80 ± 5.46	13.19
Model 3	−2.26 ± 4.51	3.93 ± 3.16	5.55	−7.72 ± 4.75	7.94 ± 4.37	10.26
Model 4	−1.64 ± 4.40	**3.65 ± 2.95**	**4.29**	**−2.50 ± 5.25**	**4.43 ± 3.75**	6.15
	Offset corrected					
	Same preprocessing			**New preprocessing**		
	Accuracy		**Robustness**	**Accuracy**		**Robustness**
	ME (SD)	MAE (SD)	mMAE	ME	MAE	mMAE
Model 1	**0.0 ± 4.21**	**3.31 ± 2.60**	4.16	**0.0 ± 4.64**	**3.71 ± 2.78**	**3.80**
Model 2	0.0 ± 5.53	4.21 ± 3.58	4.89	0.0 ± 5.84	4.68 ± 3.50	5.25
Model 3	0.0 ± 4.51	3.51 ± 2.82	4.41	0.0 ± 4.75	3.81 ± 2.84	4.03
Model 4	0.0 ± 4.40	3.45 ± 2.73	**4.09**	0.0 ± 5.25	4.07 ± 3.30	5.75

dataset (seen) and the other different (unseen) (cf. Section 4.1.1). We predicted the age using all 20 previously trained models (cf. Section 5.1) on 1493 T1w baseline scans from the UKB dataset. The predictions are showcased as scatter plots in Fig. 5.

Performance evaluation in Table 3 shows that, while all models captured the linear trend of aging, a systematic offset parallel to the identity line can be observed. All models underestimated age across the whole age interval, which was especially evident for predictions on data with unseen T1w preprocessing. The MAE of Models 1 and 4, using the same preprocessing, was equal to 3.73 and 3.65 years, and increased by less than a year when applied to T1w scans with the unseen preprocessing. This increase was much larger for the Models 2 and 3, with the MAE increasing from 4.32 and 3.93 years to almost 10 and 8 years. The difference was even more pronounced when observing mMAE, which increased to over 10 years.

Compared to the results on the multi-site test dataset (Table 2), the MAE of regression Models 1, 2, and 3 increased by about 0.75 years; however, the increase was smallest for classification Model 4, at 0.4 years.

Offset correction improved both accuracy and robustness metrics (cf. Table 3, *top vs. bottom*). Compared to results on the multi-site test dataset, the increase in MAE due to unseen site was 0.34 years, with an additional 0.45 years due to unseen T1w preprocessing. The offset-corrected metrics for the new site with both the same and new T1w preprocessing are visually summarized in the bottom-right yellow area of the radar plot in Fig. 2. Since ME equals to 0 and is the same for all models, it was not included in the plot.

Statistical evaluation involved fitting two LMEMs on the offset-corrected mean ensemble predictions: the first for predictions on the unseen dataset, either with the same or unseen T1w preprocessing. The LMEMs were fit with AE as the dependent variable, subject ID as the random effect, and model architecture as the fixed effect.

The results of the ANOVA showed that model architecture was significant for both the same ($F(3, 4476) = 55.7$, $p < 0.001$) and unseen T1w preprocessing ($F(3, 4476) = 53.9$, $p < 0.001$). The post-hoc pairwise differences between the EMMs of LMEM fit on data with the same T1w preprocessing showed a statistically significant difference between

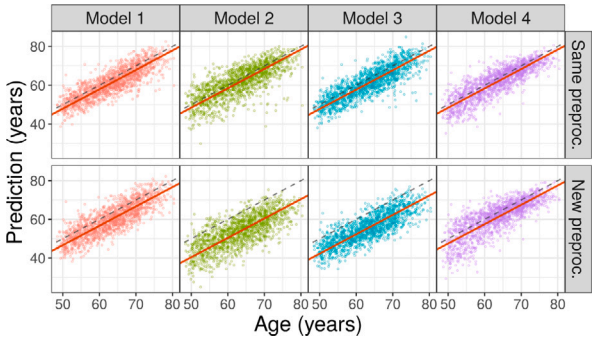

Fig. 5. Mean ensemble based age predictions on the unseen dataset with the same and unseen T1w preprocessing (*upper*) and *bottom rows*, respectively). The corrected offset is marked with the red line.

Model 2 and all other models ($p < 0.001$) and between Model 1 and 3 ($p = 0.041$). However, the post-hoc pairwise analysis on unseen T1w preprocessing data showed a statistically significant difference between all pairs ($p < 0.001$; $p = 0.009$ between Model 3 and 4), except between Models 1 and 3 ($p = 0.658$). Coefficient estimates and their 95% CI are reported in the Supplementary Table 9.

The LMEM and ANOVA analyses were also tested with the sex variable and its interaction with other variables as fixed effects. ANOVA indicated no significant differences in MAE with respect to sex ($F(1, 245) = 0.004$, $p = 0.952$), nor did it show statistical significance between the interaction of sex and model architecture ($F(3, 735) = 0.004$, $p = 0.203$) (results not shown). These findings assert that the accuracy of age predictions remains stable across sex groups.

5.3. Test-retest reproducibility

Using brain age as a biomarker necessitates consistent age predictions on MRIs taken within a short time span, having low intra-model variance, despite potential accuracy bias. To verify this, we applied all

Table 4
Reproducibility metrics for the mean ensemble and intraclass correlation (ICC) for the models trained with $K = 5$ different weight initializations. All values are in years; the best values are highlighted in **bold**.

	Reproducibility			ICC	
	$\overline{\sigma}(y'_{scan})$	$\mu(d)$	$\overline{\sigma}(d)$	$y'^{(k)}_{scan}$	$d^{(k)}$
Model 1	2.02	**−0.03**	0.86	**0.983**	0.549
Model 2	**1.97**	−0.10	0.77	**0.989**	**0.590**
Model 3	4.04	−0.09	1.01	0.945	0.455
Model 4	2.15	−0.09	**0.63**	**0.985**	0.489

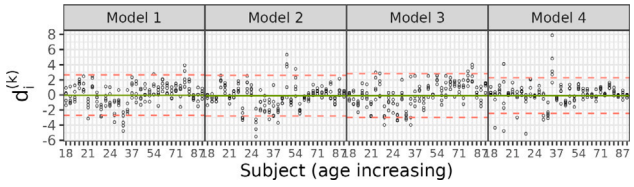

Fig. 6. Predicted age difference (*vertical axis*) between two scans for a subset of subjects (*horizontal axis*). Each point represents one of the five models, trained with different weight initializations. Subjects were arranged in ascending order of age (from left to right), with every tenth individual selected for plotting. For each model, the average predicted age difference (computed over all subjects, not just the plotted ones) is marked in *green*, and the 95% CI is indicated by *red lines*.

Table 5
Consistency metrics for the mean ensemble age predictions on the longitudinal dataset. The MAdE intervals are set based on the age at baseline. All numbers are in years, best are marked in **bold**.

	Consistency		
	MdE	MAdE	mMAdE
Model 1	**−0.52 ± 1.52**	1.20 ± 1.05	1.91
Model 2	−0.65 ± 1.43	1.15 ± 1.01	2.06
Model 3	−0.52 ± 1.61	1.38 ± 0.98	2.03
Model 4	−0.90 ± 0.92	**1.05 ± 0.74**	**1.52**

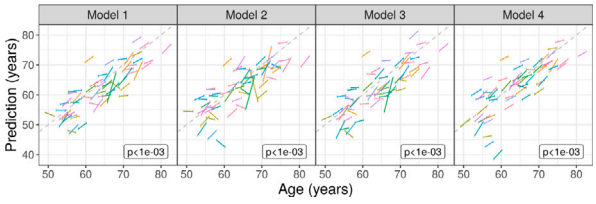

Fig. 7. Age trajectories between the baseline and the follow-up T1w scans based on the true and predicted age (for 60 randomly chosen subjects, one colored line per subject). The *p*-values reject the hypothesis, that the average slope equals to 1 (cf. text for details).

20 models resulting from the experiment described in Section 5.1 to obtain age predictions on the test-retest dataset. We then computed reproducibility metrics and conducted statistical analyses using LMEM and ANOVA.

The reproducibility results are summarized in Table 4 for five trained models ($K = 5$) and two scans ($t = 2$) per subject. The average difference between the first and second scan, $\mu(d)$, ranged from −0.03 for Model 1 to −0.10 years for Model 2. The average standard deviation of scan predictions, $\overline{\sigma}(y'_{scan})$, was lowest for Model 2 at 1.97 years, followed by Model 1 at 2.02 years.

Fig. 6 displays the age prediction difference, $d_i^{(k)}$, between the two scans for each subject. Each of the five points represents models with $K = 5$ different weight initializations. The difference in age predictions remained consistent within subjects, with values close to 0. For some subjects, the age prediction difference reached up to four years. Notably, for Model 4, there was minimal within-subject variation, indicating that the large difference in age prediction was consistent for all five models with $K = 5$ different weight initializations. As a result, the average standard deviation of differences, $\overline{\sigma}(d)$, was lowest for Model 4 (cf. Table 4), at 0.63 years.

The agreement in the predicted difference among the 5 models was computed using ICC, with Model 2 achieving the highest level of agreement with an ICC of 0.59 (cf. Table 4). However, the results showed moderate to poor reliability for all four models. Yet, the ICC for each individual T1w scan was much excellent for all models, ranging from 0.95 for Model 3 to 0.98 for Models 1, 2 and 4. We infer that the differences stem from the quality of input T1w scans, especially for the lower input resolution of Model 3, and that the models generally exhibit good reproducibility. The values of all metrics in Table 4 are visually summarized in the bottom-left red area of the radar plot in Fig. 2.

The observations above are supported by statistical analyses. Specifically, we fitted a LMEM with prediction difference $d_i^{(k)}$ as the dependent variable, subject ID as a random effect, and the model architecture as a fixed effect. Pairwise marginal means show that none of the paired differences are statistically significant ($p > 0.05$). The ANOVA test did not identify the model architecture as statistically significant ($F(3, 3531) = 2.097$, $p = 0.098$). The exact regression coefficients, their 95% CI, and ANOVA F-values are reported in Supplementary Table 10.

5.4. Longitudinal consistency

All 20 models trained on the multi-site dataset (Section 5.1), were applied on the longitudinal dataset (which had the same T1w preprocessing as the multi-site dataset). Subsequently, mean ensemble predictions were computed, and consistency metrics were evaluated. Results in Table 5 show that, generally, the MdE (i.e. ME between the actual and predicted age difference) was negative, and all models on underestimated the age difference between visits, with values ranging from 0.52 to 0.9 years. Model 4 achieved best longitudinal accuracy and robustness (lowest MAdE and mMAdE values, respectively), despite exhibiting the largest bias, as indicated by the highest MdE. The MAdE error corresponded to 50%–90% of the average age span between scans. Values of the three metrics from Table 5 are visually summarized in the top-left green area of the radar plot in Fig. 2.

Fig. 7 shows the age trajectories based on the chronological versus the predicted brain age between the baseline and follow-up visit for approximately 60 randomly chosen subjects and their corresponding T1w scans from the UKB test set. We expect to observe the slopes close to or equal to the identity line (*dashed diagonal line* in Fig. 7). For Model 2, the subject-specific age differences follow a rather randomized pattern, while for Models 1 and 4, the majority of the subject-specific lines seem parallel to the identity line.

For statistical confirmation, the ith subject slope was evaluated as $k_i = \frac{y'_{i2} - y'_{i1}}{y_{i2} - y_{i1}}$, where (y_{i1}, y'_{i1}) denote the baseline age and its estimate and (y_{i2}, y'_{i2}) denote the follow-up age and its estimate. The null hypothesis, that the average slope $\overline{k}$ across subject was different from 1, was rejected for all models. The average slope with $\overline{k} = 0.96$ was closest to 1 for Model 4.

Finally, we fit a LMEM model with MAdE as the dependent variable, model architecture as a fixed factor and subject ID as a random factor. The estimated coefficients significantly differ between architectures. Additionally, ANOVA was statistically significant ($F(3, 4476) = 109.26$, $p < 0.001$). Post-hoc pairwise analysis of EMMs showed statistical significance between all pairs ($p < 0.001$) except Model 1 and Model 4 ($p = 0.248$). Exact coefficients are presented in the Supplementary Table 11.

6. Discussion

We proposed the Brain Age Standardized Evaluation or BASE protocol and showcased a comprehensive, objective, quantitative, and

Table 6
The overall performance rank and individual task rankings for each model. Within each task, as shown in Fig. 2, models were first ranked based on each metric. Subsequently, an average rank was computed for each task by aggregating these metric ranks. For the final overall ranking, the average ranks from all tasks were consolidated, resulting in an overall model ranking.

	Multi-site Acc&Rob S 5.1	Unseen site Acc&Rob S 5.2	Test-retest Reproducibility S 5.3	Longitudinal Consistency S 5.4	Overall
Model 1	**1**	**1**	**1**	2	**1**
Model 2	4	4	2	4	4
Model 3	2	2	4	3	3
Model 4	3	2	3	**1**	2

reproducible validation and comparison of four DL-based brain age prediction models. The principal results of using BASE are visually summarized in Fig. 2.

The proposed datasets and evaluation protocol in BASE represent a framework that ensures reproducibility across different studies, as it considers and tackles the confounding factors impacting the variability of results. Namely, the use of heterogeneous, multi-site and multi-source datasets induces variability of results caused by MRI scanner-specific and biological (subject) variability, while the use of multiple T1w preprocessing pipelines induces variability of results caused by the use of specific tools and implementations. To account for model (epistemic) uncertainty, we adopted repeated model training by using five different seeds for random model weight initialization and incorporated this in a statistical framework based on LMEMs.

We introduced the BASE evaluation in conjunction with four datasets, each corresponding to a specific aspect. When provided with a suitable dataset, BASE can be applied to various other datasets, including those from other modalities such as functional and diffusion tensor MRI and positron emission tomography. However, the results from this study, as well as any other, are directly comparable only when applied to the same datasets, which are subjected to identical preprocessing procedures. Alterations in dataset attributes or variations in preprocessing can have a significant impact on model outcomes. Although model rankings based on accuracy largely remained the same when changing the preprocessing, there can be variations in MAE values, which may hinder comparisons across studies (Dular et al., 2023).

We developed a detailed set of performance metrics tailored to evaluate the accuracy, robustness, reproducibility, and consistency of brain age models. Based on research objectives, specific components of the BASE evaluation can be favored. For example, given its best ranking in longitudinal consistency (cf. Table 6) and its comparable reproducibility, Model 4 emerges as the prime choice for patient monitoring. Considering its accuracy and robustness across both known and unseen sites, Model 1 is best-suited for population studies, out of the models compared.

6.1. Accuracy and robustness

In addition to MAE, which is the main metric used in brain age estimation, we propose the inclusion of ME as a complementary measure. ME allows for the assessment of the offset across the entire age interval, which is particularly insightful when models are applied on unseen site dataset (Section 5.2). Furthermore, we recommend reporting standard deviations of MAE and ME, so as to evaluate the model precision. While many studies report the MAE along with its standard deviation, it is essential to clarify that this standard deviation is typically computed over the MAE values obtained from repeated model training with different weight initialization or cross-validation folds, rather than across all subjects. The former provides insights into model reproducibility, while the latter offers information on prediction dispersion. In this paper, we argue for and recommend reporting the latter as it provides valuable information on prediction variability.

We introduced the robustness metric mMAE, where a large discrepancy between MAE and mMAE can serve as an indicator that MAE is biased due to differences in age structure or age span between the training and test datasets. For instance, Han et al. (2022) reported and overall MAE of 3.72 years, whereas low MAE of 2.86 and 2.97 years was obtained on two large pediatric datasets (age < 22) with over 10000 subjects, but a large MAE of 5.35 years on 252 adults up to the age of 60.

While RMSE, r, and R^2 are commonly reported in brain age studies, we did not include them in BASE. Note that the values of RMSE are represented by the standard deviation of ME, and thus redundant. Furthermore, it may take up to 4 decimals of r in order to detect the difference in model performance (He et al., 2022b), whereas the proposed metrics are more sensitive to differences in performance.

6.2. Performance on unseen site dataset

Brain age models are generally applied on new (unseen) cohorts, where the anticipated goal is to estimate the brain age gap between healthy individuals and those with specific condition; consequently, our model needs to provide accurate age assessment for healthy controls.

The observed drop in performance on unseen site datasets, i.e. about 0.7 years increase in MAE, aligns with existing literature in brain age studies. For instance, Feng et al. (2020) found a minor increase in MAE of 0.15 years, Jonsson et al. (2019) a larger increase of about 3 and 5 years on two unseen datasets, and Dartora et al. (2022) an increase of 0.92 and 3.04 years on two unseen datasets for a model trained on minimally preprocessed T1w images. As our results show, the differences in the T1w preprocessing contribute to a substantial drop in performance, such as the increase of MAE above 1 year.

Ranking of models according to MAE may be relevant for best model selection, if the MAE increases are consistent. He et al. (2021) evaluated the performance of three distinct models on three unseen datasets and observed an overall increase in MAE of approximately 0.7 and up to one year. Despite changes in MAE, the rank order of models' accuracy among the three datasets remained consistent.

Our findings are mirrored, as similar performance ranks were observed on the unseen, as well as on the seen data, and even on unseen data with different T1w preprocessing. The increase of MAE was systematic across the entire age span, but varied depending on the model and dataset. This observation is apparent from Fig. 5, which shows Models 1 and 4 as less susceptible to changes in the dataset and T1w preprocessing. Furthermore, all models tend to perform better on datasets that bear resemblance to the T1w preprocessing of the training set (Dular et al., 2023).

6.2.1. Offset correction

As a result of regression dilution, researchers often observed a systematic over- and under-estimation of brain age on lower and upper end of the dataset age span. To alleviate this phenomenon, many researches apply post-hoc correction of the predictions in form of (linear) bias correction (de Lange et al., 2019; Peng et al., 2021; Cole et al., 2017a; Smith et al., 2019; Cheng et al., 2021; Dunås et al., 2021), fitting a regression line on training or validation dataset. However, recent studies (de Lange et al., 2022; Butler et al., 2021) caution against the use of such corrections, since it can inflate performance metrics.

Upon visual inspection of Fig. 5, the increase in MAE seems systematic across the whole age span and specific to the model and dataset. This systematic offset was also reported by Franke and Gaser (2012), who proposed that the increase in MAE is dataset-specific, resulting in a consistent offset in subject's age predictions across multiple scanning time points.

We propose correcting this offset when testing on new unseen site dataset, however, the offset-uncorrected model predictions should always be inspected and reported, in order to evaluate the validity of predicted brain age estimates.

The offset correction does not compromise the reproducibility and consistency metrics, while the accuracy and robustness metrics are improved. It is important to mention, that a model with poor predictive power and a significant bias towards the mean, will still yield poor performance, even after the offset correction. Unlike fitting a linear regression model, our approach does not result in overly optimistic performance.

6.3. Reproducibility

We demonstrate that the most accurate Model 1 is not necessarily the most reproducible, as can be clearly observed from Fig. 2. Specifically, Model 4 achieved the smallest average standard deviation of age prediction, as well as one of the highest values of ICC. Surprisingly, despite its poor accuracy, Model 2 exhibited the lowest average variability in age predictions for models trained with different weight initialization. Reproducibility metrics are invariant to offset by design, as the aim is to focus on a model's ability to reproduce the same prediction. Models with low variance but potentially high bias will still perform well. Thus, these metrics should be viewed as complementary to accuracy metrics, rather than a replacement for them.

The reported ICC values above 0.94, are comparable to 0.9 reported by Franke and Gaser (2012). Despite a high ICC of up to 0.99, the standard deviation of age predictions for a single MRI was at best 1.97 years, which is comparable to the 1.88 years reported by He et al. (2021). The sensitivity of model training to becoming trapped into local optima might present a significant challenge to using brain age as an individualized clinical biomarker. Employing model ensembling appears to be a promising strategy to mitigate the effects of random model weight initialization.

6.4. Consistency

The evaluation of consistency encompasses the use of baseline and follow-up T1w MRIs, assessing the accuracy and robustness of predicted age differences using the MdE, MAdE, and mMAdE metrics, analogous to ME, MAE, and mMAE metrics. Despite achieving accurate and reproducible results, we observe that all tested models often fall short when predicting age across longitudinal data. We found the mean values of slopes are statistically different from the ideal value of 1, with even the best-performing models exhibiting an average age difference error of 1.2 years, which is about half of the actual average time difference of 2.25 years.

There is a clear need to design models specifically tailored to address consistency. Incorporating longitudinal data might offer a solution, as it could enable us to model individual aging trajectories (Levakov et al., 2020). Dartora et al. (2022) used multiple images per subject in the training dataset and their visual results appear more desirable compared to results of this study. However, an objective and quantitative evaluation using the proposed consistency metrics is needed before drawing conclusions.

Given that longitudinal data are scarce, DL-based data augmentation could be leveraged. For instance, Fu et al. (2023) developed a methodology for generating missing data in longitudinal cohorts with anatomically plausible images. This approach could prove beneficial in enhancing the dataset for better model performance.

6.5. Study reproducibility: Data, code and BASE protocol

The standardized dataset comprises multi-site train, validation, and test T1w scans from 2504 healthy subjects. Additionally, there are two test sets: one with previously unseen site longitudinal T1w MRIs (N_{subj} = 1493, N_{scan} = 2986) and another with test-retest T1w MRIs from 316 subjects, ranging from 18 to 94 years of age. All T1w MRI scans used in this study were sourced from public datasets.[6] Every scan underwent a rigorous visual quality assessment to exclude low-quality scans or those with unsuccessful T1w preprocessing.

To ensure the reproducibility of our study, we have disclosed at the public GitHub repository[7] the subject ID lists, dataset split, the implementations and dependencies of the T1w preprocessing routines, brain age regression models, scripts to re-run the experiments and carry out the performance evaluations and statistical analyses. With the use of BASE implementation other researchers may evaluate novel models and techniques in a standardized manner.

Although a large public dataset for brain age dubbed OpenBHB (Dufumier et al., 2021, 2022), has recently become available, it falls short in certain critical aspects of brain age performance assessment. Specifically, the OpenBHB dataset lacks longitudinal and test-retest datasets, but which are essential for the assessment of consistency and reproducibility as per the BASE protocol. Hence, there was a need to introduce a new dataset. Moreover, the OpenBHB has biased age structure since 40% of its MRI scans are from subjects aged between 20 and 25 years, skewing the mean age to 25 years, as compared to 52 in our dataset. This huge age bias can lead to an artificially low MAE value, as the brain age predictions are generally more accurate at lower age, thereby presenting an overoptimistic and biased evaluation of the model's performance across the 18–95 years age span.

6.6. Statistical framework

Point estimates of performance metrics like the MAE, which are usually reported in brain age literature, need to be statistically evaluated to enable one to draw generalizing conclusions. For this purpose we used the LMEMs, as they enable to account for repeated measures on a subject level by including the subject ID as a random effect. Our results show that despite the observed difference in MAE point estimates the difference may not be statistically significant. For instance, when comparing the performances of Models 1 and 4 (cf. Table 2), the seemingly relevant difference in MAE values of about 0.3 years was not statistically significant (cf. Fig. 4).

6.7. Study limitation

In this study, we have concentrated our efforts on a select group of four CNN-based models, each showcasing significant variations in terms of input dimensionality, image resolution, and output representation. While this selection enables a clear and focused introduction of BASE, providing insights into its operation across different models and application scenarios, we acknowledge that it does not cover the exhaustive array of available model architectures, including various branches of convolutional networks and emerging transformer architectures. While a broader comparison could potentially yield a more comprehensive understanding of the BASE approach, our intention was to introduce BASE with clarity and precision, demonstrating its applicability. We encourage future work in this area to apply BASE, either partially for their specific application, or in whole, across a broader spectrum of models.

[6] While some public datasets necessitate online registration to access the T1w MRI scans, the UKB dataset requires a fee.

[7] GitHub repository at https://github.com/AralRalud/BASE.git

7. Conclusion

In this study we proposed and demonstrated the application of the Brain Age Standard Evaluation or BASE. The BASE comprises the dataset, performance metrics and an evaluation protocol. Using BASE we evaluated four state-of-the-art deep regression brain age models in aspects such as accuracy and robustness on multi-site and unseen site and differently preprocessed T1w MRIs, reproducibility on test-retest and consistency on longitudinal T1w scans. Our study is fully reproducible as the dataset information and code are made publicly available at https://github.com/AralRalud/BASE.git.

CRediT authorship contribution statement

Lara Dular: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Žiga Špiclin:** Conceptualization, Methodology, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

None

Data availability

Data used in the study was obtained from public data sources. Some require online registration to gain access to the MRI scans, acquiring the UK Biobank dataset necessitates a fee payment.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT-4 in order to improve the readability of this paper. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

Data collection and sharing for this project was partially provided by:

- **Cambridge Centre for Ageing and Neuroscience (CamCAN).** CamCAN funding was provided by the UK Biotechnology and Biological Sciences Research Council (grant number BB/H008217/1), together with support from the UK Medical Research Council and University of Cambridge, UK.
- **OASIS-1: Cross-Sectional:** Principal Investigators: D. Marcus, R, Buckner, J, Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382
- **OASIS-2: Longitudinal:** Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.
- **ABIDE I.** Primary support for the work by Adriana Di Martino was provided by the (NIMH K23MH087770) and the Leon Levy Foundation.
  Primary support for the work by Michael P. Milham and the INDI team was provided by gifts from Joseph P. Healy and the Stavros Niarchos Foundation to the Child Mind Institute, as well as by an NIMH award to MPM (NIMH R03MH096321).
- **UK Biobank Resource** under Application Number 68981.
- **Alzheimer's Disease Neuroimaging Initiative (ADNI).** 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.

Appendix A. Supplementary materials

A.1. Brain age model training

A.1.1. Dataset split

The multi-site dataset was split into training, validation and test set, assuring comparable age distribution across all three set. The age distribution of multi-site dataset, as well as unseen site dataset and longitudinal dataset is shown in Fig. 8.

A.1.2. Loss function and optimization

Loss function was selected based on the task formulation as a regression or classification problem. Models 1, 2 and 3 were trained using L1 loss, while Model 4 the Kullback–Leibler divergence (KLD). For regression models MSE was also tested, however it often diverged and resulted in worse overall performance (results not shown).

We used the SGD algorithm with momentum 0.9 as proposed in three out of four studies (Cole et al., 2017b; Peng et al., 2021; Ueda et al., 2019), keeping the learning rate (LR) decay schedule as originally proposed for each individual model. Namely, the LR decay schedule proposed for Model 1 was to decay the initial LR for 3% after each epoch, for Model 4 we multiplied the LR by 0.3 every 30 epochs. For Model 2 and 3 we computed the LR on ith epoch as $LR_i = \frac{LR_0}{1+(i\lambda)}$, where LR_0 denotes the starting LR and λ the learning rate decay.

We experimentally determined that Models 1 and 4 typically converged after 110 epochs, while Model 2 and 3 converged after 400 epochs.

A.1.3. Hyperparameter tuning

We used the multi-site validation dataset to determine hyperparameter values for each of the four models. The LR and batch size hyperparameter values for each model were chosen based on a wide grid search, which was set around the proposed values in corresponding original papers. For instance, tested LR values were 10^{-2}, 10^{-3}, 10^{-4}, $5 \cdot 10^{-5}$, 10^{-5}, and 10^{-6}. The batch size for Models 2 and 3 was set to 4, 8, 16, 32 and 64. Due to GPU constraints we trained Model

Table 7
Proposed hyperparameter values in original literature and the values implemented herein. Only the hyperparameters marked with * were reevaluated. The resulting model accuracy is reported as MAE in years.

	Model 1		Model 2	
	Proposed	**Implemented**	**Proposed**	**Implemented**
Input size	$182 \times 218 \times 182$	$157 \times 189 \times 170$	$157 \times 189 \times 15$	
*Batch size	28	16	16	32
*Loss function	L1		MSE	L1
*Learning rate (LR)	1×10^{-2}	1×10^{-4}	1×10^{-4}	1×10^{-3}
LR decay	3%		1×10^{-4}	
Weight decay	5×10^{-5}		1×10^{-3}	
Momentum	0.9		0.9	
Parameters	≈ 900 000		≈ 6.6 mio	
MAE (Test) [years][a] *med*[*min*, *max*]	4.65	3.57 [3.52, 3.61]	4.0	4.23 [4.14, 4.67]
	Model 3		**Model 4**	
	Proposed	**Implemented**	**Proposed**	**Implemented**
Input size	$95 \times 79 \times 78$		$160 \times 192 \times 160$	$157 \times 189 \times 170$
*Batch size	16	8	8	8
*Loss function	MSE	L1	KLD	
*Learning rate (LR)	5×10^{-5}		1×10^{-2}	
LR decay	1×10^{-4}		×0.3 every 30 epochs	
Weight decay	5×10^{-4}		1×10^{-3}	
Momentum	0.9		0.9	
Parameters	≈ 900 000		≈ 6.6 mio	
MAE (Test) [years][a] *med*[*min*, *max*]	3.67	3.57 [3.52, 4.26]	2.14	3.35 [3.29, 3.42]

[a] Test MAE of implemented models is presented as the median, minimal and maximal values of the last 10 epochs of model training.

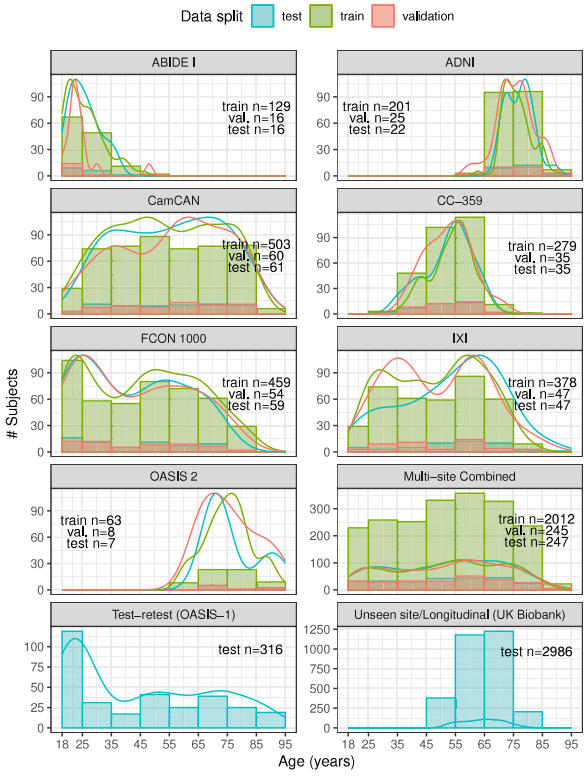

Fig. 8. Density of age distribution per each and combined multi-site dataset, depicted for train, test and validation set splits.

1 with batch size 4, 8, 16 and 24 and Model 4 with batch size 4 and 8. Hyperparameter values were selected based on their associated model performance, which was evaluated using the median value of MAE on the validation set across the last 10 epochs. The chosen hyperparameters are presented in Supplementary Table 7.

A.1.4. Data augmentation

In all our experiments models were trained using the following data augmentation: *(1)* random shifting along all major axes with probability of 0.3 for an integer sampled from $[-s, s]$, where $s = 3$ for Model 3 and $s = 5$ for Models 1,2, and 4; *(2)* random padding with probability of 0.3 for an integer from range $[0, p]$, where $p = 2$ for Model 3 and $p = 5$ for Models 1,2, and 4; *(3)* flipping over central sagittal plane with probability of 0.5. Note that the difference in size of parameters s and p for Model 3 in comparison to the other three models is due to difference in input sizes between models as the result of downsampling.

The image size for 2D Model 2 and Model 3, trained on downsampled images was adapted during augmentation. Namely, with Model 2 the 15 axial slices (predefined in atlas space) were sampled to obtain input image size of $157 \times 189 \times 15$, while with Model 3 the input images were downsampled using sinc resampling and cropped to size $95 \times 79 \times 78$.

A.1.5. Weighted training

Weighted training is a strategy of assigning higher sampling probabilities to subjects in underrepresented age categories, such that the expected number of samples from each age category becomes equal. Since Model 4 is defined as a classification model, it is susceptible to higher error in underrepresented age classes. The use of weighted training improved the predictions of Model 4 on the multi-site validation set for age > 80 years, but not for the other three models, which was confirmed by LME model (results not shown).

Specifically, we applied weighted random sampler with replacement during training, assigning each subject a weight of N/n_i, where n_i denotes the number of samples in category i. Subjects were split into age categories [18, 20), [20, 25), [25, 30), …, [85, 90), [90, 100) as previously proposed by Feng et al. (2020). The number of sampled subjects was set to N to keep the number of samples per train epoch the same as in the experiments without weighted training.

A.2. Detailed results of statistical analyses

In the following Tables 8–11 we show detailed results of the ANOVA test and LMEM as performed in respective Sections 5.1–5.4. The levels

Table 8
Results of LMEM and ANOVA for evaluating the impact of model architecture (Section 5.1) for mean ensemble strategy on multi-site dataset, with absolute error as response variable, model as fixed factor and subject ID as random factor: $|y' - y| = Model + (1|ID)$.

Multi-site data							
ANOVA			**LMEM**	Estimate	Std. Error	2.5%	97.5%
F value	7.709	**	Intercept	2.960	0.170	2.626	3.295
NumDF	3		Model 2	0.757	0.162	0.439	1.075
DenDF	738		Model 3	0.217	0.162	−0.101	0.535
			Model 4	0.289	0.162	−0.029	0.607
			Random effects	Variance	Std.Dev.		
			Subject ID (Intercept)	3.3939	1.985		
			Residual	3.326	1.805		

Table 9
Results of LMEM and ANOVA of ensemble model performance on new site dataset (Section 5.2) with same and different preprocessing than the one used in model training, with absolute difference error as response variable, model architecture as fixed factor and subject ID as random factor: $|y' - y| = Model + (1|ID)$.

Same preprocessing							
ANOVA			**LMEM**	Estimate	Std. Error	2.5%	97.5%
F value	55.70	***	Intercept	3.313	0.076	3.163	3.463
NumDF	3		Model 2	0.899	0.076	0.749	1.048
DenDF	4476		Model 3	0.201	0.076	0.052	0.351
			Model 4	0.134	0.076	−0.015	0.284
			Random effects	Variance	Std.Dev.		
			Subject ID (Intercept)	4.410	2.100		
			Residual	4.335	2.082		
New preprocessing							
ANOVA			**LMEM**	Estimate	Std. Error	2.5%	97.5%
F value	53.87	***	Intercept	3.714	0.081	3.555	3.872
NumDF	3		Model 2	0.966	0.084	0.802	1.130
DenDF	4476		Model 3	0.096	0.084	−0.068	0.260
			Model 4	0.170	0.360	0.196	0.524
			Random effects	Variance	Std.Dev.		
			Subject ID (Intercept)	4.496	2.120		
			Residual	5.232	2.287		

Table 10
Results of LMEM and ANOVA evaluating the four models on test-retest dataset (Section 5.3). The predicted difference serves as the response variable, with model architecture as the fixed factor and subject ID as the random factor: $d = Model + (1|ID)$.

Test-retest dataset						
ANOVA		**LMEM**	Estimate	Std. Error	2.5%	97.5%
F value	1.13	Intercept	−0.032	0.051	−0.132	0.067
NumDF	3	Model 2	−0.069	0.040	−0.149	0.011
DenDF	6001	Model 3	−0.056	0.040	−0.136	0.023
		Model 4	−0.054	0.040	−0.134	0.025
		Random effects	Variance	Std.Dev.		
		Subject ID (Intercept)	0.546	0.739		
		Residual	1.308	1.144		

Table 11
Results of LMEM and ANOVA on same site longitudinal dataset and new site longitudinal dataset (Section 5.4), with absolute difference error as response variable, model architecture as fixed factor and subject ID as random factor: $AdE = Model + (1|ID)$.

Same site longitudinal data						
ANOVA		**LMEM**	Estimate	Std. Error	2.5%	97.5%
F value	2.43	Intercept	1.209	0.124	0.967	1.456
NumDF	3	Model 2	−0.056	0.126	−0.304	0.191
DenDF	325.64	Model 3	0.177	0.126	−0.070	0.425
		Model 4	−0.155	0.126	−0.403	0.093
		Random effects	Variance	Std.Dev.		
		Subject ID (Intercept)	0.206	0.454		
		Residual	0.728	0.853		

(*continued on next page*)

Table 11 (*continued*).

New site longitudinal data							
ANOVA			**LMEM**	Estimate	Std. Error	2.5%	97.5%
F value	109.26	***	Intercept	1.294	0.057	1.183	1.405
NumDF	3		Model 2	0.811	0.056	0.701	0.921
DenDF	4476		Model 3	0.372	0.056	0.262	0.482
			Model 4	−0.104	0.056	−0.214	−0.006
			Random effects	Variance	Std.Dev.		
			Subject ID (Intercept)	2.435	1.560		
			Residual	2.352	1.534		

of statistical significance are denoted as: '***' for $0 < p < 0.001$, '**' for $0.001 < p < 0.01$, '*' for $0.01 < p < 0.05$ and '•' for $0.05 < p < 0.1$.

References

Amoroso, N., La Rocca, M., Bellantuono, L., Diacono, D., Fanizzi, A., Lella, E., Lombardi, A., Maggipinto, T., Monaco, A., Tangaro, S., Bellotti, R., 2019. Deep learning and multiplex networks for accurate modeling of brain age. Front. Aging Neurosci. 11, http://dx.doi.org/10.3389/fnagi.2019.00115, URL https://www.frontiersin.org/articles/10.3389/fnagi.2019.00115/full.

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

Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C., Mechelli, A., 2021b. Machine learning for brain age prediction: Introduction to methods and clinical applications. eBioMedicine 72, http://dx.doi.org/10.1016/j.ebiom.2021.103600, URL https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(21)00393-5/fulltext.

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., Wolf, D., Gur, R.E., Gur, R.C., Morris, J., Albert, M.S., Grabe, H.J., Resnick, S., Bryan, R.N., Wolk, D.A., Shou, H., Davatzikos, C., 2020. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14468 individuals worldwide. Brain 143 (7), 2312–2324. http://dx.doi.org/10.1093/brain/awaa160.

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 Inf. 26 (4), 1432–1440. http://dx.doi.org/10.1109/JBHI.2021.3083187, URL https://ieeexplore.ieee.org/document/9439893.

Beheshti, I., Potvin, O., Duchesne, S., 2021. Patch-wise brain age longitudinal reliability. Hum. Brain Mapp. 42 (3), 690–698. http://dx.doi.org/10.1002/hbm.25253, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25253.

Bellantuono, L., Marzano, L., La Rocca, M., Duncan, D., Lombardi, A., Maggipinto, T., Monaco, A., Tangaro, S., Amoroso, N., Bellotti, R., 2021. Predicting brain age with complex networks: From adolescence to adulthood. NeuroImage 225, 117458. http://dx.doi.org/10.1016/j.neuroimage.2020.117458, URL https://www.sciencedirect.com/science/article/pii/S1053811920309435.

Butler, E.R., Chen, A., Ramadan, R., Le, T.T., Ruparel, K., Moore, T.M., Satterthwaite, T.D., Zhang, F., Shou, H., Gur, R.C., Nichols, T.E., Shinohara, R.T., 2021. Pitfalls in brain age analyses. Hum. Brain Mapp. 42 (13), 4092–4101. http://dx.doi.org/10.1002/hbm.25533, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25533.

Cai, H., Gao, Y., Liu, M., 2023. Graph transformer geometric learning of brain networks using multimodal MR images for brain age estimation. IEEE Trans. Med. Imaging 42 (2), 456–466. http://dx.doi.org/10.1109/TMI.2022.3222093.

Cheng, J., Liu, Z., Guan, H., Wu, Z., Zhu, H., Jiang, J., Wen, W., Tao, D., Liua, T., 2021. Brain age estimation from MRI using cascade networks with ranking loss. IEEE Trans. Med. Imaging 1. http://dx.doi.org/10.1109/TMI.2021.3085948.

Cole, J.H., Annus, T., Wilson, L.R., Remtulla, R., Hong, Y.T., Fryer, T.D., Acosta-Cabronero, J., Cardenas-Blanco, A., Smith, R., Menon, D.K., Zaman, S.H., Nestor, P.J., Holland, A.J., 2017a. Brain-predicted age in down syndrome is associated with beta amyloid deposition and cognitive decline. Neurobiol. Aging 56, 41–49. http://dx.doi.org/10.1016/j.neurobiolaging.2017.04.006, URL http://www.sciencedirect.com/science/article/pii/S0197458017301240.

Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G., 2017b. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124. http://dx.doi.org/10.1016/j.neuroimage.2017.07.059, URL http://www.sciencedirect.com/science/article/pii/S1053811917306407.

Cole, J.H., Raffel, J., Friede, T., Eshaghi, A., Brownlee, W.J., Chard, D., Stefano, N.D., Enzinger, C., Pirpamer, L., Filippi, M., Gasperini, C., Rocca, M.A., Rovira, A., Ruggieri, S., Sastre-Garriga, J., Stromillo, M.L., Uitdehaag, B.M.J., Vrenken, H., Barkhof, F., Nicholas, R., Ciccarelli, O., 2020. Longitudinal assessment of multiple sclerosis with the brain-age paradigm. Ann. Neurol. 88 (1), 93–105. http://dx.doi.org/10.1002/ana.25746, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25746.

Cole, J.H., Underwood, J., Caan, M.W.A., Francesco, D.D., Zoest, R.A.v., Leech, R., Wit, F.W.N.M., Portegies, P., Geurtsen, G.J., Schmand, B.A., Loeff, M.F.S.v.d., Franceschi, C., Sabin, C.A., Majoie, C.B.L.M., Winston, A., Reiss, P., Sharp, D.J., 2017c. Increased brain-predicted aging in treated HIV disease. Neurology 88 (14), 1349–1357. http://dx.doi.org/10.1212/WNL.0000000000003790, URL https://n.neurology.org/content/88/14/1349.

Couvy-Duchesne, B., Faouzi, J., Martin, B., Thibeau-Sutre, E., Wild, A., Ansart, M., Durrleman, S., Dormont, D., Burgos, N., Colliot, O., 2020. Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: ARAMIS contribution to the predictive analytics competition 2019 challenge. Front. Psychiatry 11, URL https://www.frontiersin.org/article/10.3389/fpsyt.2020.593336.

Dartora, C., Marseglia, A., Mårtensson, G., Rukh, G., Dang, J., Muehlboeck, J.-S., Wahlund, L.-O., Moreno, R., Barroso, J., Ferreira, D., Schiöth, H.B., Westman, E., 2022. Predicting the age of the brain with minimally processed T1-weighted MRI data. http://dx.doi.org/10.1101/2022.09.06.22279594, medRxiv, URL https://doi.org/10.1101/2022.09.06.22279594.

de Lange, A.-M.G., Anatürk, M., Rokicki, J., Han, L.K.M., Franke, K., Alnæs, D., Ebmeier, K.P., Draganski, B., Kaufmann, T., Westlye, L.T., Hahn, T., Cole, J.H., 2022. Mind the gap: Performance metric evaluation in brain-age prediction. Hum. Brain Mapp. 43 (10), 3113–3129. http://dx.doi.org/10.1002/hbm.25837, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25837.

de Lange, A.-M.G., Kaufmann, T., van der Meer, D., Maglanoc, L.A., Alnæs, D., Moberget, T., Douaud, G., Andreassen, O.A., Westlye, L.T., 2019. Population-based neuroimaging reveals traces of childbirth in the maternal brain. Proc. Natl. Acad. Sci. USA 116 (44), 22341–22346. http://dx.doi.org/10.1073/pnas.1910666116, URL https://www.pnas.org/content/116/44/22341.

Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E., 2021. OpenBHB: a Multi-Site Brain MRI Dataset for Age Prediction and Debiasing. IEEE Dataport, http://dx.doi.org/10.21227/7jsg-jx57.

Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E., 2022. OpenBHB: a large-scale multi-site brain MRI data-set for age prediction and debiasing. Neuroimage 263, 119637. http://dx.doi.org/10.1016/j.neuroimage.2022.119637.

Dular, L., Pernus, F., Spiclin, Z., 2023. Extensive T1-weighted MRI preprocessing improves generalizability of deep brain age prediction models. http://dx.doi.org/10.1101/2023.05.10.540134, bioRxiv, arXiv:https://www.biorxiv.org/content/early/2023/10/30/2023.05.10.540134.full.pdf.

Dunås, T., Wåhlin, A., Nyberg, L., Boraxbekk, C.-J., 2021. Multimodal image analysis of apparent brain age identifies physical fitness as predictor of brain maintenance. Cerebral Cortex (bhab019), http://dx.doi.org/10.1093/cercor/bhab019.

Feng, X., Lipton, Z.C., Yang, J., Small, S.A., Provenzano, F.A., 2020. Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging. Neurobiol. Aging 91, 15–25. http://dx.doi.org/10.1016/j.neurobiolaging.2020.02.009, URL https://www.sciencedirect.com/science/article/pii/S0197458020300361.

Finn, R., 1970. A note on estimating the reliability of categorical data. Edu. Psychol. Measur. 30 (1), 71–76. http://dx.doi.org/10.1177/001316447003000106.

Fonov, V., Evans, A., McKinstry, R., Almli, C., Collins, D., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Organization for Human Brain Mapping 2009 Annual Meeting, NeuroImage Organization for Human Brain Mapping 2009 Annual Meeting, 47, S102. http://dx.doi.org/10.1016/S1053-8119(09)70884-5.URL https://www.sciencedirect.com/science/article/pii/S1053811909708845,

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. http://dx.doi.org/10.1024/1662-9647/a000074, URL https://econtent.hogrefe.com/doi/10.1024/1662-9647/a000074.

Franke, K., Gaser, C., Manor, B., Novak, V., 2013. Advanced BrainAGE in older adults with type 2 diabetes mellitus. Front. Aging Neurosci. 5, 90. http://dx.doi.org/10.3389/fnagi.2013.00090.

Fu, J., Tzortzakakis, A., Barroso, J., Westman, E., Ferreira, D., Moreno, R., Initiative, for the Alzheimer's Disease Neuroimaging, 2023. Fast three-dimensional image generation for healthy brain aging using diffeomorphic registration. Hum. Brain

Mapp. 44 (4), 1289–1308. http://dx.doi.org/10.1002/hbm.26165, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.26165.

Han, J., Kim, S.Y., Lee, J., Lee, W.H., 2022. Brain age prediction: A comparison between machine learning models using brain morphometric data. Sensors 22 (20), 8077. http://dx.doi.org/10.3390/s22208077, URL https://www.mdpi.com/1424-8220/22/20/8077.

He, S., Feng, Y., Grant, P.E., Ou, Y., 2022a. Deep relation learning for regression and its application to brain age estimation. IEEE Trans. Med. Imaging 41 (9), 2304–2317. http://dx.doi.org/10.1109/TMI.2022.3161739.

He, S., Grant, P.E., Ou, Y., 2022b. Global-local transformer for brain age estimation. IEEE Trans. Med. Imaging 41 (1), 213–224. http://dx.doi.org/10.1109/TMI.2021.3108910.

He, S., Pereira, D., David Perez, J., Gollub, R.L., Murphy, S.N., Prabhu, S., Pienaar, R., Robertson, R.L., Ellen Grant, P., Ou, Y., 2021. Multi-channel attention-fusion neural network for brain age estimation: Accuracy, generality, and interpretation with 16,705 healthy MRIs across lifespan. Med. Image Anal. 72, 102091. http://dx.doi.org/10.1016/j.media.2021.102091, URL https://www.sciencedirect.com/science/article/pii/S1361841521001377.

Høgestøl, E.A., Kaufmann, T., Nygaard, G.O., Beyer, M.K., Sowa, P., Nordvik, J.E., Kolskår, K., Richard, G., Andreassen, O.A., Harbo, H.F., Westlye, L.T., 2019. Cross-sectional and longitudinal MRI brain scans reveal accelerated brain aging in multiple sclerosis. Front Neurol 10, 450. http://dx.doi.org/10.3389/fneur.2019.00450.

Huang, T., Chen, H., Fujimoto, R., Ito, K., Wu, K., Sato, K., Taki, Y., Fukuda, H., Aoki, T., 2017. Age estimation from brain MRI images using deep learning. In: 2017 IEEE 14th International Symposium on Biomedical Imaging. ISBI 2017, pp. 849–852. http://dx.doi.org/10.1109/ISBI.2017.7950650.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841. http://dx.doi.org/10.1016/s1053-8119(02)91132-8.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156. http://dx.doi.org/10.1016/s1361-8415(01)00036-6.

Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., Ellingsen, L.M., Walters, G.B., Gudbjartsson, D.F., Stefansson, H., Stefansson, K., Ulfarsson, M.O., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nature Commun. 10 (1), 5409. http://dx.doi.org/10.1038/s41467-019-13163-9, URL https://www.nature.com/articles/s41467-019-13163-9.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., Pantelis, C., Meisenzahl, E., 2014. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophr. Bull. 40 (5), 1140–1153. http://dx.doi.org/10.1093/schbul/sbt142.

Kuo, C.-Y., Tai, T.-M., Lee, P.-L., Tseng, C.-W., Chen, C.-Y., Chen, L.-K., Lee, C.-K., Chou, K.-H., See, S., Lin, C.-P., 2021. Improving individual brain age prediction using an ensemble deep learning framework. Front. Psychiatry 12, http://dx.doi.org/10.3389/fpsyt.2021.626677, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8021919/.

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. http://dx.doi.org/10.1002/hbm.25011, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25011.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. J. Magn. Reson. Imaging 31 (1), 192–203. http://dx.doi.org/10.1002/jmri.22003.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J. Cogn. Neurosci. 22 (12), 2677–2684. http://dx.doi.org/10.1162/jocn.2009.21407.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): Cross-sectional MRI data in Young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507. http://dx.doi.org/10.1162/jocn.2007.19.9.1498, arXiv:https://direct.mit.edu/jocn/article-pdf/19/9/1498/1936514/jocn.2007.19.9.1498.pdf.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19 (11), 1523–1536. http://dx.doi.org/10.1038/nn.4393, URL https://www.nature.com/articles/nn.4393.

Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., Ourselin, S., 2014. Global image registration using a symmetric block-matching approach. J. Med. Imaging 1 (2), 1–6. http://dx.doi.org/10.1117/1.JMI.1.2.024003.

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. http://dx.doi.org/10.1016/j.neuroimage.2023.119947, URL https://www.sciencedirect.com/science/article/pii/S1053811923000940.

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. http://dx.doi.org/10.1016/j.media.2020.101871, URL http://www.sciencedirect.com/science/article/pii/S1361841520302358.

Petersen, K.J., Metcalf, N., Cooley, S., Tomov, D., Vaida, F., Paul, R., Ances, B.M., 2021. Accelerated brain aging and cerebral blood flow reduction in persons with human immunodeficiency virus. Clin. Infect. Dis. (ciab169), http://dx.doi.org/10.1093/cid/ciab169.

Ronan, L., Alexander-Bloch, A.F., Wagstyl, K., Farooqi, S., Brayne, C., Tyler, L.K., Fletcher, P.C., 2016. Obesity associated with increased brain age from midlife. Neurobiol. Aging 47, 63–70. http://dx.doi.org/10.1016/j.neurobiolaging.2016.07.010, URL http://www.sciencedirect.com/science/article/pii/S0197458016301403.

Schnack, H.G., van Haren, N.E., Nieuwenhuis, M., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., 2016. Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study. AJP 173 (6), 607–616. http://dx.doi.org/10.1176/appi.ajp.2015.15070922, URL https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2015.15070922.

Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., Calder, A.J., Marslen-Wilson, W.D., Duncan, J., Dalgleish, T., Henson, R.N., Brayne, C., Matthews, F.E., 2014. The cambridge centre for ageing and neuroscience (cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurol. 14, http://dx.doi.org/10.1186/s12883-014-0204-1, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4219118/.

Smith, S.M., Alfaro-Almagro, F., Miller, K.L., 2022. UK biobank brain imaging documentation. URL https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf.

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. http://dx.doi.org/10.1016/j.neuroimage.2019.06.017, URL http://www.sciencedirect.com/science/article/pii/S1053811919305026.

Souza, R., Lucena, O., Garrafa, J., Gobbi, D., Saluzzi, M., Appenzeller, S., Rittner, L., Frayne, R., Lotufo, R., 2018. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. NeuroImage 170, 482–494. http://dx.doi.org/10.1016/j.neuroimage.2017.08.021, URL http://www.sciencedirect.com/science/article/pii/S1053811917306687.

Tanveer, M., Ganaie, M., Beheshti, I., Goel, T., Ahmad, N., Lai, K.-T., Huang, K., Zhang, Y.-D., Del Ser, J., Lin, C.-T., 2023. Deep learning for brain age estimation: A systematic review. Inf. Fusion 96, 130–143. http://dx.doi.org/10.1016/j.inffus.2023.03.007, URL https://www.sciencedirect.com/science/article/pii/S156625352300088X.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-Can, n., Henson, R.N., 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. http://dx.doi.org/10.1016/j.neuroimage.2015.09.018.

Terock, J., Bonk, S., Frenzel, S., Wittfeld, K., Garvert, L., Hosten, N., Nauck, M., Völzke, H., Auwera, S.V.d., Grabe, H.J., 2022. Vitamin d deficit is associated with accelerated brain aging in the general population. Psychiatry Res. Neuroimag. 327, 111558. http://dx.doi.org/10.1016/j.pscychresns.2022.111558, URL https://www.sciencedirect.com/science/article/pii/S0925492722001172.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. http://dx.doi.org/10.1109/TMI.2010.2046908.

Ueda, M., Ito, K., Wu, K., Sato, K., Taki, Y., Fukuda, H., Aoki, T., 2019. An age estimation method using 3D-CNN from brain MRI images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI2019, pp. 380–383. http://dx.doi.org/10.1109/ISBI.2019.8759392.

Xiong, M., Lin, L., Jin, Y., Kang, W., Wu, S., Sun, S., 2023. Comparison of machine learning models for brain age prediction using six imaging modalities on middle-aged and older adults. Sensors 23 (7), 3622. http://dx.doi.org/10.3390/s23073622, URL https://www.mdpi.com/1424-8220/23/7/3622.