Computers in Biology and Medicine 173 (2024) 108320

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Extensive T1-weighted MRI preprocessing improves generalizability of deep brain age prediction models☆

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

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

ARTICLE INFO

Keywords:
Brain age
MRI preprocessing
Deep regression models
Linear mixed effect models
Dataset bias
Transfer learning
Reproducible research
UK Biobank

ABSTRACT

Brain age is an estimate of chronological age obtained from T1-weighted magnetic resonance images (T1w MRI), representing a straightforward diagnostic biomarker of brain aging and associated diseases. While the current best accuracy of brain age predictions on T1w MRIs of healthy subjects ranges from two to three years, comparing results across studies is challenging due to differences in the datasets, T1w preprocessing pipelines, and evaluation protocols used. This paper investigates the impact of T1w image preprocessing on the performance of four deep learning brain age models from recent literature. Four preprocessing pipelines, which differed in terms of registration transform, grayscale correction, and software implementation, were evaluated. The results showed that the choice of software or preprocessing steps could significantly affect the prediction error, with a maximum increase of 0.75 years in mean absolute error (MAE) for the same model and dataset. While grayscale correction had no significant impact on MAE, using affine rather than rigid registration to brain atlas statistically significantly improved MAE. Models trained on 3D images with isotropic 1 mm^3 resolution exhibited less sensitivity to the T1w preprocessing variations compared to 2D models or those trained on downsampled 3D images. Our findings indicate that extensive T1w preprocessing improves MAE, especially when predicting on a new dataset. This runs counter to prevailing research literature, which suggests that models trained on minimally preprocessed T1w scans are better suited for age predictions on MRIs from unseen scanners. We demonstrate that, irrespective of the model or T1w preprocessing used during training, applying some form of offset correction is essential to enable the model's performance to generalize effectively on datasets from unseen sites, regardless of whether they have undergone the same or different T1w preprocessing as the training set.

1. Introduction

Brain age, a neurological biomarker of individual brain health [1], has emerged as a pivotal measure of biological aging over the last decade. By measuring the discrepancy between brain age and chronological age, premature brain aging has been demonstrated in neurological diseases and disorders such as Alzheimer's dementia [2], Multiple Sclerosis [3,4], and other diseases,including Type 2 Diabetes [5], and Human Immunodeficiency Virus (HIV) infection [6,7]. Brain age may deviate from chronological age even for healthy individuals, with research indicating its association with various environmental and lifestyle factors such as tobacco and alcohol consumption [5,8–10]. Evaluating the age gap therefore represents an evolving diagnostic biomarker, opening an avenue for researchers to uncover patterns and heterogeneity in the aging process.

Deep learning (DL) methods, particularly convolutional neural networks (CNNs), have revolutionized the training of brain age algorithms by enabling direct processing of medical images, circumventing the need for prior feature extraction. These algorithms have been effectively employed across various neuroimaging modalities, with the majority involving T1-weighted (T1w) magnetic resonance imaging (MRI), while T2-weighted [10–12], T2-FLAIR [10], diffusion tensor imaging (DTI) [13,14], functional MRI [10,11,15], and PET [16] have also been used. Due to a large number of publicly available T1w brain MRI

☆ This document is the results of the research project funded by the Slovenian Research and Innovation Agency (Core Research Grant No. P2-0232 and Research Grants Nos. J2-2500 and J2-3059).

* 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.compbiomed.2024.108320
Received 23 February 2023; Received in revised form 9 January 2024; Accepted 12 March 2024
Available online 20 March 2024

datasets, this study focused on the impact of T1w MRI preprocessing on the performance of brain age regression models.

The shift towards deep learning in this field marks a significant departure from earlier traditional machine learning models, enhancing the precision of brain age models. However, due to differences in T1w MRI preprocessing pipelines and software implementations, it is difficult to disentangle the contribution of methodological innovations from the impact of the T1w preprocessing. The preprocessing pipelines used in brain age studies generally include gray scale enhancement, such as bias field corrections [17–20], and registration to a brain atlas. Degrees of freedom in registration of T1w to atlas space also varies from rigid [21,22] to general linear (affine) [17–20,23,24] or even nonlinear transforms [18,25,26]. Skull stripping that involves extracting the brain from surrounding tissues was also applied in certain studies [17,19,20,25–27]. A comprehensive study on natural images found that the effect of image preprocessing and augmentation on the performance of regression models was greater than the effects of variability in model architecture [28], highlighting the need to analyze the effect of T1w preprocessing approach for achieving best accuracy and reproducibility of brain age prediction models.

Besides training on the T1w MRIs, brain age models are often trained on Gray Matter (GM) and White Matter (WM) segmentation maps. Studies show that models trained on the former far outperform models trained on the latter [18,22], with reports of models trained on GM even outperforming models trained on preprocessed T1w MRIs [19,29]. However, related neuroimaging studies show that measurements of cortical surface thickness differ significantly depending on the pipeline applied [30,31] and reveal a significant discrepancy in cortical thickness reproducibility metrics [32]. Reasons could also include T1w MRI resolution variations and contrast-to-noise differences. It seems that using GM segmentation for brain age predictions is rather ill-posed and, therefore, this study will focus on preprocessed T1w images as model input. However, differences in T1w preprocessing may also arise from the use of different software implementations [27]. It is yet to be determined if there is a significant effect of the software implementations on brain age prediction, even for a fairly simple T1w preprocessing approaches.

For clinical application, it is crucial to validate brain age prediction models on datasets from new, previously unseen sites that were not used during model training. Such scenario simulates the real-world application of applying a pretrained model to new data, which may have undergone different preprocessing. While Feng et al. [20] reported a rather small increase in Mean Absolute Error (MAE) of 0.15 years, multiple other deep learning studies have shown a larger deterioration in accuracy, with increases in MAE ranging from one to five years [21,27,33]. Interestingly, even when the same pretrained model and preprocessing method were used, the deterioration in accuracy on different new sites could vary substantially [21,33]. This suggests that the observed increase in error is more likely due to the unique characteristics of the new dataset and/or its preprocessing, rather than a lack of generalizability in the model itself.

To address the aforementioned challenges and uncertainties in the field of brain age prediction, particularly the varying impact of T1w MRI preprocessing, software implementations, and the generalizability of models across diverse datasets, this study aims to provide a systematic and detailed analysis with the following contributions:

(i) A thorough and reproducible quantitative evaluation of the impact of four different T1w preprocessing pipelines on the accuracy of brain age prediction. These pipelines differ in the level of preprocessing and the software tools employed.
(ii) A rigorous statistical evaluation involving repeated model training with random initialization and use of linear mixed-effects models (LMEMs), encompassing the study of the impact of various confounding factors.
(iii) Study of model performance generalization on an unseen site dataset and/or new T1w preprocessing pipeline and software implementation.

2. Materials and methods

2.1. Datasets

For studying the effect of image preprocessing on brain age prediction, we created two datasets: (i) a multi-site dataset for training, validation and testing and (ii) a new unseen site dataset used solely for testing. All included subjects were healthy individuals, without previously known neurological diseases, from 18 to 95 years old.

The **multi-site dataset** comprises T1 W MRIs, gathered from seven publicly available datasets, totaling 4428 T1w MRIs of healthy subjects. Most datasets within this collection sourced images from multiple hospitals or sites, utilizing an array of MRI scanners, including those from vendors GE, Siemens, and Philips, operating at 1.5T and 3T field strengths. Exceptions are the OASIS 2 and CamCAN datasets, in which all scans were acquired on a single scanner. Due to the integration of these multi-source, multi-site, and multi-vendor datasets, variations in acquisition protocols are inherently present.

The T1w images scans in the multi-site dataset were preprocessed using four different preprocessing pipelines, described in Section 2.2, and underwent a visual quality control. Images that did not pass the visual quality control for reasons like motion artifacts, failed preprocessing, etc., were excluded ($N_{excl} = 408$). Furthermore, subjects under the age of 18 or with missing age information were discarded ($N_{disc} = 481$) and, in case multiple scans per subject were available, a single scan (chronologically the first non-discarded image) was retained. Finally accepted were a total of 2504 T1w MRIs, which were split into train ($N = 2012$), validation ($N = 245$) and test ($N = 247$) datasets. The descriptive statistics per dataset are given in Supplementary Table 4. For reproducibility reasons, the exact subject IDs included in each split are provided on our code repository.

The **unseen site dataset** was a subset of the UK Biobank (UKB) dataset and included 1493 T1w MRI scans of healthy subjects. All included subjects met the inclusion criteria of not having long-standing illnesses and were required to self-report an overall health rating of *excellent* or *good* at the time of scan acquisition. The raw defaced T1w MRIs were preprocessed using the same four preprocessing pipelines as the multi-site dataset, and all scans passed the visual quality control. In addition, a fifth preprocessing pipeline was already applied to the dataset by the UKB dataset providers.

The ground truth brain age corresponds to the subject's chronological age, which was either given by the dataset providers or calculated from the provided date of birth and the MRI acquisition date. For the majority of datasets, including ADNI, CamCAN, CC-359, OASIS 2, and FCON 1000, the age was provided as a rounded figure to the nearest year. The age distribution of the included T1w subject scans per dataset, and the train/validation/test subsets, is provided in Supplementary Materials (Table 4, Fig. 7).

2.2. Image preprocessing pipelines

We implemented four preprocessing pipelines using a combination of publicly available and in-house developed software. These pipelines vary with respect to the spatial transform applied in the registration to atlas space, the extent of gray scale corrections, as well as the algorithm and software implementations used. For clarity, a schematic representation of the four pipelines is illustrated in Fig. 1.

Common to all four pipelines, the input T1w image was first converted to the Nifti format. In the first three pipelines, the input raw T1w image was initially denoised using Adaptive non-local means denoising[2] with spatially varying noise levels [34].

[2] Adaptive non-local means denoising implementation: https://github.com/djkwon/naonlm3d.

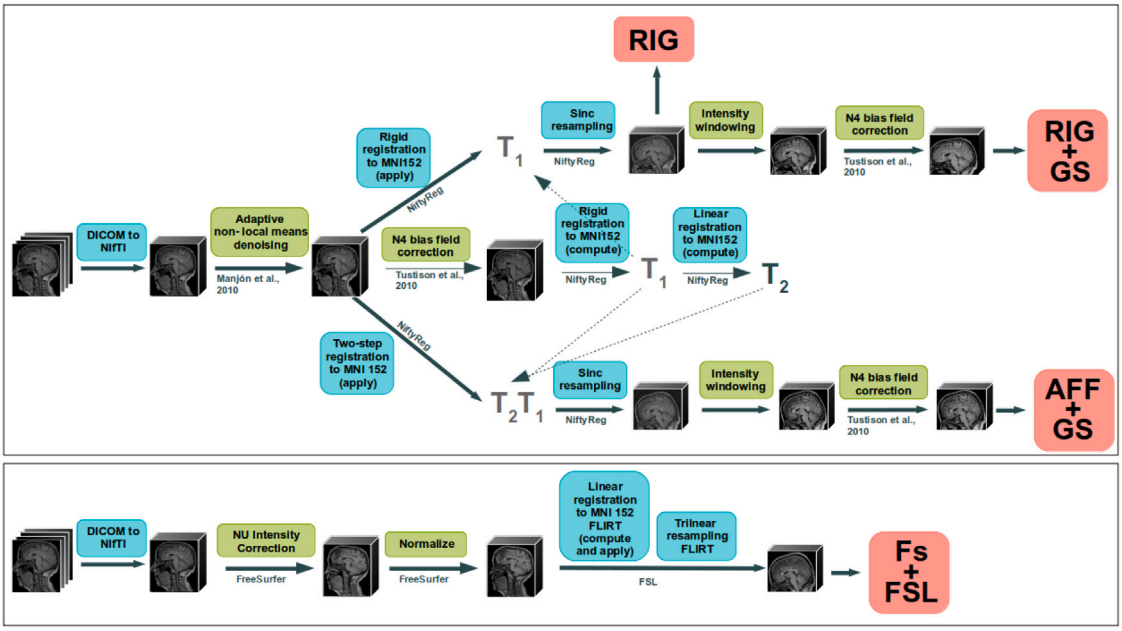


Fig. 1. Schematic representation of preprocessing pipelines and software used.

Aligned with Cole et al. [22], the first pipeline, denoted **RIG**, performed rigid registration of the denoised T1w image into the MNI152 nonlinear atlas, version 2009c [35] (7th generation), with size 193 × 229 × 193 and spacing 1 mm^3. To improve registration accuracy, intensity inhomogeneity correction (without mask) was applied to the denoised image using N4 algorithm[3] [36], prior to running the registration. The inhomogeneity corrected T1w image was used during registration only, while, finally, the denoised T1w image was sinc resampled using the obtained rigid transformation.

The second pipeline, **RIG+GS**, extended the RIG pipeline by applying an additional two-step grayscale correction procedure to the RIG output. The first step, *(1)* intensity windowing, involves computation of the lower and upper thresholds based on the grayscale histogram, smoothed with a Gaussian filter. The lower threshold is set based on 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 evaluating 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.

The third pipeline, **AFF+GS**, was a modified version of the RIG+GS, by applying in sequence the rigid and affine registration steps. Finally, the two-step grayscale correction procedure was applied as in the RIG+GS pipeline. All previously mentioned image registration and resampling steps for processing pipelines RIG, RIG+GS and AFF+GS were performed using the publicly available NiftyReg software[4] [37].

The fourth pipeline, **Fs+FSL**, utilized commonly used software tools FreeSurfer[5] and FSL (FMRIB Software Library)[6] [38] and included gray scale corrections and affine registration. Raw T1w images were preprocessed using the grayscale correction preprocessing stages of FreeSurfer's cortical reconstruction `recon-all` pipeline, with default parameter settings. The preprocessing entails non-parametric non-uniform intensity normalization (N3), followed by intensity normalization that sets the mean intensity of the white matter to 110 [39]. In order to ensure consistency among all preprocessing pipelines, we also applied registration to the MNI152 nonlinear atlas, version 2009c [35], the same reference space as used in previous pipelines. Specifically, we used FSL FLIRT [40] with default settings, performing linear registration with trilinear resampling.

For testing on the unseen site dataset, an additional fifth **UKB preprocessing** variant was exclusively applied to the UKB dataset. This preprocessing, already executed by the dataset providers, is detailed by Smith et al. [41]. Each preprocessed T1w image was resampled to the MNI152 nonlinear 6th generation atlas [42] using FSL FLIRT [40, 43]. Both the preprocessed MRI and the linear transformation matrix for this process were provided by the UKB. Considering the alignment with the 7th generation MNI152 atlas of our four implemented preprocessing pipelines, we applied an additional common linear registration between 6th and 7th generation atlas spaces, we applied an additional common linear registration between 6th and 7th generation atlas spaces, followed by 3rd order resampling. This step ensured spatial alignment across all preprocessed MRIs. The linear transformation matrix between the two MNI spaces was pre-computed using FSL FLIRT.

2.3. Age prediction models

To study the effect of preprocessing in relation to model architecture, four fundamentally different CNN models for brain age estimation were reimplemented based on the descriptions in the literature. Only minor alterations, such as adjustments for the input image dimensions, were made to assure comparability across the experiments.

Model 1, proposed by Cole et al. [22], was a convolutional CNN trained on full resolution 3D T1w MRIs. **Model 2**, proposed by Huang et al. [24], was trained on 2D images by taking 15 equidistantly sampled axial slices as input channels. **Model 3**, proposed by Ueda et al. [23], was trained on downsampled T1w MRIs. Finally, **Model 4**, proposed by Peng et al. [18], was a fully convolutional model trained on full resolution 3D images. The architectures of the four models are depicted in Fig. 2.

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

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

[5] Freesurfer: https://surfer.nmr.mgh.harvard.edu/.

[6] FSL (FMRIB Software Library): https://fsl.fmrib.ox.ac.uk/fsl/fslwiki.

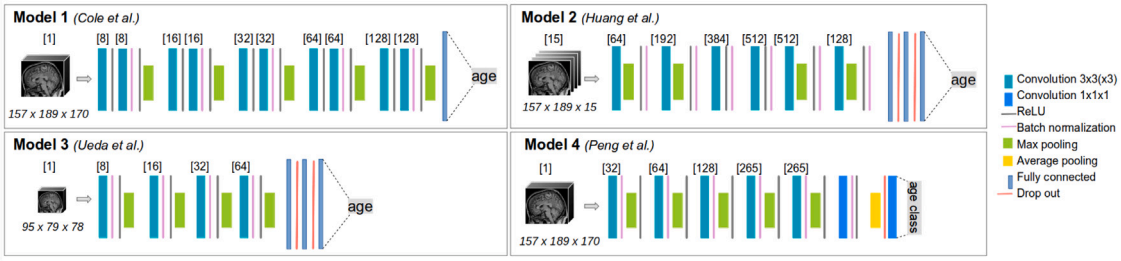


Fig. 2. Architecture of the four reimplemented CNN models for brain age prediction.

Brain age estimation is typically formulated as a regression task, such that the model outputs a non-negative real number reflecting the age of the subject based on their T1w MRI scan. Models 1, 2, and 3 therefore had linear activation in the last fully connected layer so as to output the scalar value representing the predicted age.

By contrast, Model 4 was designed as a classification model. Here, the ground truth age value y for each sample was transformed into a so-called *soft label*, represented as Gaussian probability density with mode located at the true age and unit variance. The probability density was discretized into non-overlapping 2-year age intervals by integrating the density over each age interval. The output age prediction was computed as weighted sum over the class probabilities, i.e. $y' = \sum_j p_j age_j$, where p_j denotes the probability of class j and age_j the center of the age class interval.

All models were implemented in PyTorch 1.4.0 for Python 3.6.8.

Hyperparameter tuning. The learning rate 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 learning rate 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 graphics processing unit (GPU) constraints we trained Model 1 with batch size 4, 8, 16 and 24 and Model 4 with batch size 4 and 8. All tested hyperparameter combinations and their results are given in Supplementary Fig. 8.

Hyperparameter selection was based on determining the epoch at model convergence, i.e. by observing the course of the loss function, and by observing MAE on the train and validation set in the last 10 epochs. To assure a robust choice of the hyperparameters with respect to both MAE and convergence, we computed median MAE across last 10 training epochs, and the hyperparameter values with smallest median MAE value were chosen as the optimal values.

The chosen optimal hyperparameter values in our study and the originally proposed hyperparameter values are given in Supplementary Table 5. Unless noted otherwise, we used these hyperparameters in all subsequent experiments.

Loss function. The choice of loss function depended on the model formulation as either regression or classification network. For Models 1, 2 and 3, we tested mean squared error (MSE) and L1 losses for multiple hyperparameter values. Due to overall better performance and stability of training, the three models were trained with the L1 loss. Model 4, defined as a classification model, was trained with Kullback–Leibler divergence as the loss function.

Optimizer. We used the stochastic gradient descent algorithm with momentum 0.9 as proposed in three out of four studies [18,22], keeping the learning rate decay schedule as originally proposed for each individual model. We have experimentally determined that Models 1 and 4 typically converged after 110 epochs, while Model 2 and 3 converged after 400 epochs.

Data augmentation. All models were trained with the following data augmentation procedures: *(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 (downsampled 3D input T1w) 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 padding and shifting parameters are lower for Model 3, due to the larger image spacing, which is as a result of image downsampling.

Further, the image size as input to the models was adapted during the augmentation. We first we removed the non-informative empty space around the head by cropping to size $157 \times 189 \times 170$ about the image center. Further, for Model 2 the 15 axial slices (predefined in atlas space) were sampled to obtain input image size of $157 \times 189 \times 15$, while for Model 3 the input images were downsampled using sinc resampling and cropped to size $95 \times 79 \times 78$.

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. Due to the smaller number of subjects in age groups above 80, weighted training was beneficial for classification Model 4, but not for the other three models.[7]

Each subject was assigned a weight of N/n_i, where n_i denotes the number of samples in category i. Age categories were set to [18, 20), [20, 25), [25, 30), …, [85, 90), [90, 100) as previously proposed by Feng et al. [20] and sampled with replacement. The number of sampled subjects was kept equal to the number of subjects N, so that the number per training epoch was kept equal to the experiments without weighted training.

2.4. Postprocessing

Model ensembling. To avoid reporting the results of a single (possibly biased) run, each model was trained five times, with different weight initialization. The final prediction of a brain age was obtained as the average of the five model predictions with different weight initialization. On the multi-site T1w train set we trained a total of 80 models: 4× image preprocessing pipelines, 4× model architectures, and 5× random weight initialization.

Offset correction. We implemented the offset adjustment by subtracting the value of mean error (ME) from the ensemble prediction, determined as follows:

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

The ME was computed for each model/preprocessing combination. Offset correction was applied only when predicting on unseen site dataset.

[7] The application of weighted training led to a statistically significant reduction in absolute error for subjects over the age of 80 years, where the number of training samples is lower. This significance was exclusively observed for Model 4 ($p < 0.001$), whereas other models did not show such an effect (results not shown).

	Section 3.1	Section 3.2	Section 3.3
Training dataset	Multi-site dataset	Multi-site dataset	Multi-site dataset
Test dataset	Multi-site dataset	Unseen site dataset	Unseen site dataset
Test dataset preprocessing	Same as training	Same as training	New preprocessing

Fig. 3. Overview of the tested brain age train and test scenarios.

2.5. Evaluation protocol

For experiment evaluation we computed commonly used performance metrics to highlight specific aspects of the prediction model performances.

An established metric of model accuracy is the mean absolute error (MAE):

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

where y_i denotes the true age and y_i' the predicted age of ith subject. We also report mean error (ME):

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

since values of ME deviating from zero show that a model on average either under- or over-estimates age on the whole age interval. Assuming the prediction error is normally distributed around zero, we expect ME to be zero.

2.5.1. Statistical analysis

Linear mixed-effects models (LMEMs) were used to describe the relationship between a prediction's absolute error as dependent variable and response variables that were set for each research question. Each LMEM included model architecture, preprocessing procedure and their interaction as fixed effect and subject ID as random effect, such that all responses for a specific subject were shifted by a subject-specific additive value. By modeling subject ID as random effect, we account for dependent data that arises from multiple brain age predictions for the same subject under multiple conditions (preprocessing procedure, model architecture, offset correction).

We employed a stepwise approach in fitting LMEMs. Namely, the models were first constructed with the fixed factors and, subsequently, we incrementally incorporated fixed-factor interactions to increase model complexity. To evaluate the benefit of increasing model complexity, we utilized Analysis of Variance (ANOVA) for model comparison, to test if the increase in complexity resulted in a statistically significant improvement in explaining the observed variability in the data.

For the final LMEMs, we reported regression coefficients and their 95% confidence intervals, provided in Supplementary materials. Results of LMEM analyses were supported by the ANOVA test declaring statistical significance for $p < 0.01$. Throughout the manuscript, we will use the term *significant* to refer to *statistically significant* result. Further, if the main fixed factor showed a difference in responses, a post-hoc pairwise test was conducted, with confidence level of 0.95, and multiplicity adjustments using Tukey's correction.

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.

3. Experiments and results

The impact of T1w MRI image preprocessing on the accuracy and reproducibility of brain age predictions using the four CNN models was studied in three scenarios shown in Fig. 3: *(1)* tested on the same-source dataset and preprocessing as used during model training (Section 3.1), *(2)* tested on an unseen site dataset, but preprocessed in the same way as the training dataset (Section 3.2), *(3)* tested on an unseen site dataset, preprocessed differently than the training dataset (Section 3.3).

3.1. Effect of image preprocessing

Our goal is to evaluate the impact of the particular choice of image preprocessing on various CNN architectures, as described in respective Sections 2.2 and 2.3. On the multi-site T1w training set, we trained a total of 80 models: 4× image preprocessing pipelines, 4× model architectures, and 5× random weight initializations. Brain age predictions were obtained as the average age prediction of five models trained with different random weight initializations. The model accuracy metrics are presented in Table 1.

We further fit a LMEM model with model architecture and preprocessing procedure as main effects and subject ID as random effect. The ANOVA test and 95% CI interval values showed both fixed factors as significant ($F(3, 3699) = 49.49, p < 2.2e{-}16$ for model architecture; $F(3, 3699) = 5.09, p = 0.002$ for preprocessing). We increased the LMEM complexity by including the interaction of the fixed factors; however the interaction term was not significant ($F(9, 3690) = 1.14, p = 0.328$). However, due to significance of the fixed factors ($F(3, 3690) = 49.51, p < 2.2e{-}16$ for model architecture; $F(3, 3690) = 5.09, p = 0.002$ for preprocessing) and the theoretically meaningful interaction, the interaction was included in the final model despite not being significant. The LMEM coefficients, their 95% CI, and ANOVA F-values are reported in Supplementary Table 7.

The results of the LMEM post-hoc pairwise analysis are shown in Fig. 4. Model 1 outperformed the other models across all four preprocessing pipelines (cf. Table 1); however, these differences were only significant between Model 1 and Model 2 (cf. Fig. 4). Furthermore, the absolute error of Model 2 was found to be significantly higher than the MAE of all 3D models ($p < 0.001$), for all but the AFF+GS preprocessing. Out of the four, Model 2 exhibited the largest bias, as measured by ME, for the RIG, RIG+GS and AFF+GS preprocessing pipelines and the highest variability, indicated by a higher standard deviation (cf. Table 1). It only achieved MAE below 4 years when trained on the AFF+GS dataset. Notably, with this preprocessing, the performance difference between Model 2 and the other models was not significant for most pairs, as depicted in Fig. 4, but generally indicates poorer performance of Model 2 compared to the 3D counterparts.

In terms of sensitivity to applied preprocessing, Model 2's performance varied the most. Conversely, the 3D models demonstrated a more stable performance, with none of the three showing an increase in MAE beyond 0.32 years. Particularly, the Model 4 showed notable robustness to the change in applied preprocessing, registering a fluctuation in MAE within the range of 0.07 years (cf. Table 1).

Table 1
Multi-site test set results for 16 combinations of the four preprocessing pipelines and four model architectures. Best MAE values wrt. model architecture (*rows*) are marked in **bold**, while best values wrt. image preprocessing procedure (*columns*) are underlined. All numbers are in years.

	RIG		RIG+GS		AFF+GS		Fs+FSL	
	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)
Model 1	−0.46 ± 4.10	3.18 ± 2.61	−0.25 ± 4.14	3.12 ± 2.73	−0.03 ± 3.86	**2.96 ± 2.47**	0.26 ± 4.30	3.22 ± 2.86
Model 2	−0.96 ± 5.75	4.47 ± 3.73	−1.11 ± 5.49	4.29 ± 3.59	−0.80 ± 4.81	**3.72 ± 3.14**	0.13 ± 5.44	4.08 ± 3.58
Model 3	−0.13 ± 4.41	3.45 ± 2.75	−1.03 ± 4.40	3.50 ± 2.85	−0.68 ± 3.90	**3.18 ± 2.35**	−0.26 ± 4.61	3.45 ± 3.06
Model 4	−0.85 ± 4.33	3.32 ± 2.90	−0.74 ± 4.34	3.29 ± 2.92	−0.46 ± 4.21	**3.25 ± 2.70**	−0.13 ± 4.52	3.31 ± 3.07

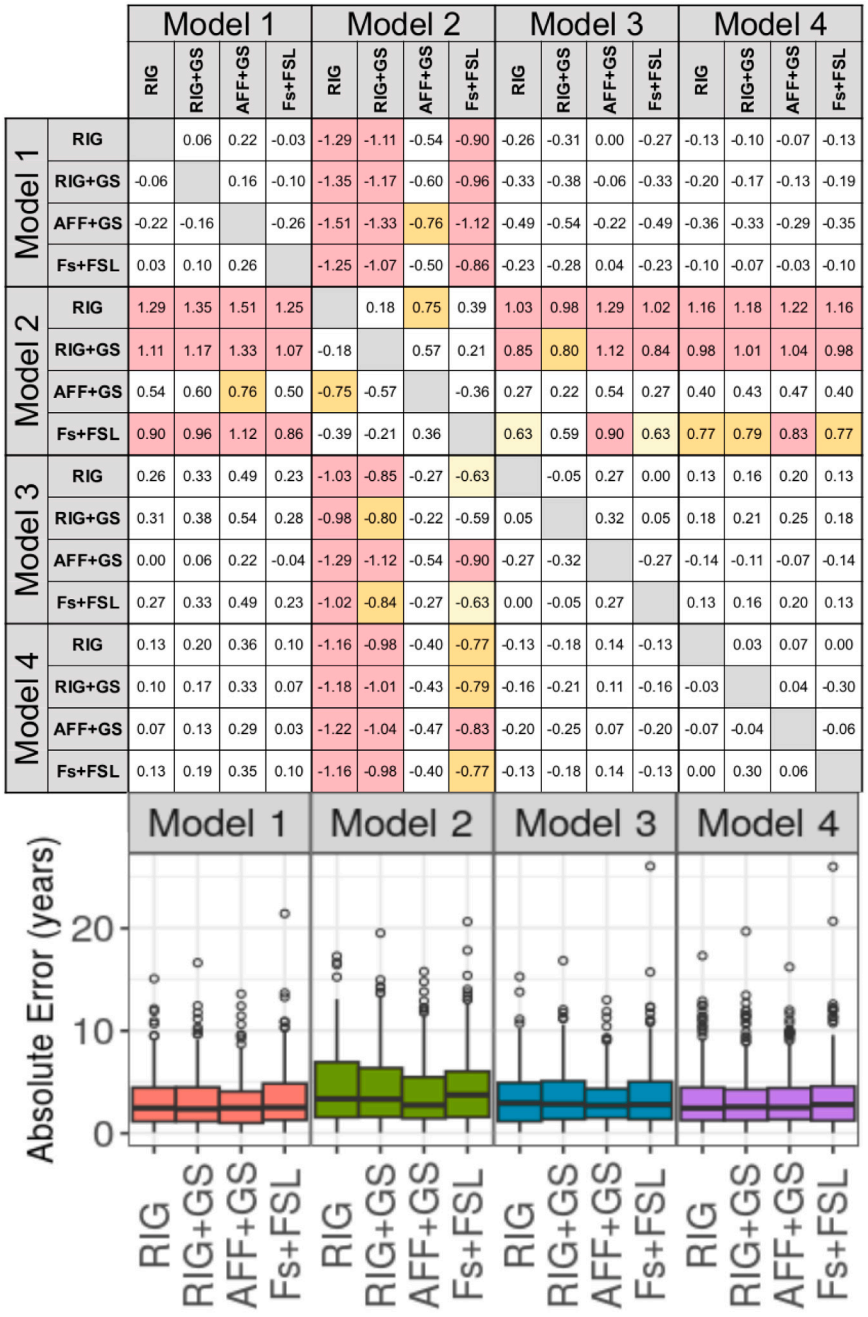


Fig. 4. Results of LMEM post-hoc pairwise statistical tests on multi-site dataset for all MRI preprocessing and model architecture combinations. The color of each square marks statistical significance: red for $p < 0.001$, orange for $0.01 \leq p < 0.001$, yellow for $0.05 \leq p < 0.01$ and white for $p > 0.05$ (not significant).

Table 2
Mean ensemble MAE values on the unseen site dataset (UKB), preprocessed in **the same** way as the multi-site training data. Results are presented for 16 model and preprocessing combinations, with (*offset*) and without (*none*) offset correction. Best MAE values wrt. model architecture (*rows*) are underlined, while best values wrt. image preprocessing procedure (*columns*) are marked in **bold**. All numbers are in years.

	Corr.	RIG		RIG+GS		AFF+GS		Fs+FSL	
		ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)
Model 1	none	−3.65 ± 4.13	4.48 ± 3.22	−3.60 ± 4.19	4.47 ± 3.25	−2.10 ± 4.21	3.73 ± 2.88	−0.73 ± 4.09	**3.33 ± 2.48**
	offset	0.0 ± 4.13	**3.26 ± 2.54**	−0.0 ± 4.19	3.33 ± 2.55	−0.0 ± 4.21	3.31 ± 2.60	−0.0 ± 4.09	3.28 ± 2.43
Model 2	none	−1.70 ± 6.26	5.07 ± 4.05	−1.65 ± 6.08	4.90 ± 3.94	−1.58 ± 5.53	**4.32 ± 3.78**	0.03 ± 5.74	4.54 ± 3.52
	offset	0.0 ± 6.26	4.95 ± 3.84	−0.0 ± 6.08	4.75 ± 3.79	0.0 ± 5.53	**4.21 ± 3.58**	−0.0 ± 5.74	4.54 ± 3.52
Model 3	none	−3.64 ± 4.56	4.73 ± 3.42	−4.38 ± 4.64	5.26 ± 3.61	−2.26 ± 4.51	3.93 ± 3.16	−2.00 ± 4.32	**3.78 ± 2.90**
	offset	−0.0 ± 4.56	3.61 ± 2.79	0.0 ± 4.64	3.70 ± 2.79	−0.0 ± 4.51	3.51 ± 2.82	−0.0 ± 4.32	**3.42 ± 2.64**
Model 4	none	−4.42 ± 4.69	5.19 ± 3.83	−3.29 ± 4.69	4.49 ± 3.55	−1.64 ± 4.4	3.65 ± 2.95	−0.71 ± 4.57	**3.63 ± 2.86**
	offset	0.0 ± 4.69	3.69 ± 2.90	−0.0 ± 4.69	3.71 ± 2.86	0.0 ± 4.40	**3.45 ± 2.73**	0.0 ± 4.57	3.59 ± 2.83

Compared to the RIG pipeline, the RIG+GS included gray scale correction steps, but resulted in only a marginal overall decrease in MAE. None of the differences were significant (cf. Fig. 4). When switching between rigid (RIG+GS) and affine registration (AFF+GS), an improvement in performance was observed for all models. Model 1 on the AFF+GS dataset achieved an MAE of 2.96 years, which was the best result reported in this study. While all models generally exhibited enhanced performance with the AFF+GS preprocessing pipeline, this improvement was significant only for Model 2 when comparing AFF+GS to RIG ($p = 0.003$).

3.2. Performance on an unseen site dataset

In this experiment, we evaluated the performance of 16 model ensembles on unseen site dataset. Generally, new data may come from a different MRI scanner or have undergone different preprocessing than the data used to train the models. We preprocessed the unseen site dataset in the same way as the training data, which is a common scenario in practice.

To evaluate the performance of the models on an unseen site dataset, we predicted brain age for all 1493 T1w scans of the UKB dataset without any additional training. For each model and preprocessing combination, we averaged the results across five pretrained models with different weight initializations. This resulted in a total of 16 predictions, one for each model and preprocessing combination, serving as our baseline.

Upon inspecting the Supplementary Fig. 9, a systematic offset in age prediction across the entire age span was observed. This offset, inherent to each combination of architecture and preprocessing prediction, can be reduced by applying an offset correction (cf. Section 2.4). The evaluation included a comparative analysis of *(1)* the baseline predictions of uncorrected mean ensemble, and *(2)* the offset-corrected predictions, computed by subtracting the ME from the predicted brain age value.

To estimate the influence of the preprocessing pipeline on model performance on the unseen dataset, a LMEM model was fitted with architecture, preprocessing, presence or absence of offset correction, their two-way and three-way interactions as fixed effects, and subject ID as a random effect. The ANOVA test confirmed that this model explained more variability than the model with no interactions ($p < 0.001$) and the model with only two-way interactions ($p < 0.001$). The ANOVA test of effects showed all main effects, their two-way, and three-way interactions as significant ($p < 0.001$). Detailed results of the LMEM model and ANOVA test are presented in the Supplementary Table 8.

The prediction errors of pretrained models on an unseen dataset are presented in Table 2. The baseline MAE values ranged from 5.26 years for the 2D Model 2 with RIG+GS preprocessing to 3.33 years for Model 1 with Fs+FSL preprocessing. Bias, as measured by the ME, demonstrates that all models on average underestimate brain age when used on an unseen dataset. Notably, the 2D model consistently exhibited the smallest bias; however, it also had the largest standard deviation in error across all preprocessing pipelines, approximately by one year.

Fig. 5 illustrates the pairwise difference in marginal means and their statistical significance between the preprocessing procedures, conditional on the model architecture and the presence or absence of offset correction, for the aforementioned LMEM. Prior to offset correction, datasets with affine correction consistently demonstrated the best performance. Specifically, for Model 1, the results from the Fs+FSL dataset outperformed those from all other preprocessing pipelines. For Models 2, 3, and 4, both Fs+FSL and AFF+GS datasets showed superior performance compared to the rest of the preprocessing pipelines, yet no significant distinction was observed between the two.

When comparing model architectures, Models 1 and 4 consistently surpassed Models 2 and 3, which were trained on reduced information. The only exception was with the RIG preprocessing procedure, where Model 1 alone excelled. Even after offset correction, the 3D models maintained a performance advantage over the 2D model.

Applying offset correction reduced the MAE by 0.54 years on average. The distinction in performance between RIG and RIG+GS remained non-significant for all models (cf. Fig. 5). Model 1 yielded the best performance across all preprocessing pipelines, achieving an overall best MAE of 3.31 ± 2.60 with the RIG preprocessing. Although the superior results from RIG might be surprising, it is critical to note that Model 1 demonstrated robustness to change in preprocessing, exhibiting MAE within the range of 0.07 years. Model 2 was the most sensitive to preprocessing. It performed best when trained with affine registration, indicating its sensitivity to spatial information.

3.3. Performance on unseen dataset with new image preprocessing

We further considered the cumulative effect of an unseen site dataset, not seen during model training, additionally with different image preprocessing as applied on dataset used for model training. The UKB was preprocessed by the dataset provider, as described in Section 2.2. Without additional training, we predicted the age for all 80 trained models. The model predictions were averaged across five models with different weight initialization, resulting in 16 predictions for each T1w MRI (baseline). As in the previous experiment, we comparatively evaluated *(i)* the baseline predictions and *(ii)* the offset-corrected predictions (cf. Section 2.4). The prediction offset on the unseen dataset is constant across the entire age span, as evidenced by Supplementary Fig. 9.

The MAE and ME metrics of the 16 mean ensembles are presented in Table 3. To estimate the influence of preprocessing pipeline, we fitted a LMEM model with architecture, preprocessing, presence or absence of offset correction, their two-way and three-way interactions as fixed effects, and subject ID as random effect. The ANOVA test confirmed that this model explained more variability than simpler LMEM models ($p < 0.001$). The ANOVA test of effects shows all main effects, their two-way, and three-way interactions as significant ($p < 0.001$). Details are provided in the Supplementary Table 9.

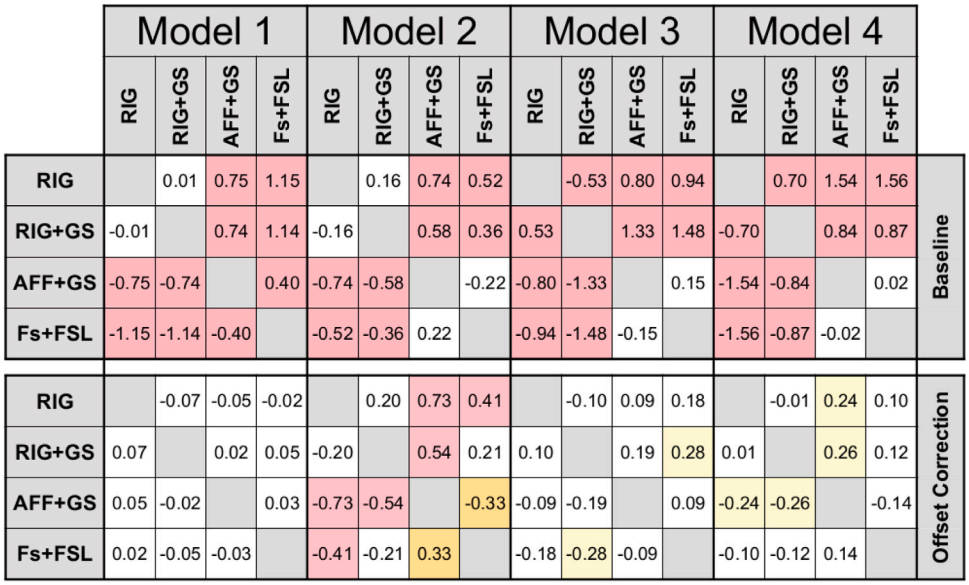


Fig. 5. The pairwise differences in marginal means between preprocessing procedures conditional on the model architecture and the presence (*lower*) or absence (*upper*) of offset correction on the unseen site dataset. The preprocessing procedure of the dataset was **the same** as the preprocessing procedure applied to the training data. The color of each square marks the significance level of difference: red for $p < 0.001$, orange for $0.01 \leq p < 0.001$, yellow for $0.05 \leq p < 0.01$ and white for $p > 0.05$ (not significant).

Table 3
Mean ensemble MAE values on the unseen site dataset (UKB), preprocessed using **a different** pipeline than used for the multi-site training data. Results are presented for 16 model and preprocessing combinations, with (*offset*) and without (*none*) additional offset correction. Best MAE values wrt. model architecture (*rows*) are underlined, while best values wrt. image preprocessing procedure (*columns*) are marked in **bold**. All numbers are in years.

	Corr.	RIG		RIG+GS		AFF+GS		Fs+FSL	
		ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)
Model 1	none	−9.14 ± 4.54	9.23 ± 4.37	−10.47 ± 4.68	10.52 ± 4.57	−3.33 ± 4.64	4.65 ± 3.31	−2.53 ± 4.55	**4.22 ± 3.04**
	offset	−0.0 ± 4.54	**3.64 ± 2.72**	−0.0 ± 4.68	3.71 ± 2.86	0.0 ± 4.64	3.71 ± 2.78	0.0 ± 4.55	3.66 ± 2.70
Model 2	none	−12.74 ± 7.02	12.91 ± 6.72	−18.20 ± 6.63	18.20 ± 6.63	−9.58 ± 5.84	9.80 ± 5.46	−7.97 ± 7.28	**8.88 ± 6.12**
	offset	−0.0 ± 7.02	5.58 ± 4.26	−0.0 ± 6.63	5.25 ± 4.05	0.0 ± 5.84	**4.68 ± 3.50**	−0.0 ± 7.28	5.90 ± 4.26
Model 3	none	−8.41 ± 4.51	8.53 ± 4.26	−14.44 ± 4.87	14.46 ± 4.82	−7.72 ± 4.75	7.94 ± 4.37	−2.82 ± 4.75	**4.45 ± 3.28**
	offset	−0.0 ± 4.51	**3.59 ± 2.72**	0.0 ± 4.87	3.88 ± 2.94	0.0 ± 4.75	3.81 ± 2.84	0.0 ± 4.75	3.75 ± 2.92
Model 4	none	−6.93 ± 5.43	7.38 ± 4.80	−11.71 ± 6.16	11.8 ± 5.98	−2.50 ± 5.25	4.43 ± 3.75	−0.06 ± 4.90	**3.83 ± 3.05**
	offset	−0.0 ± 5.43	4.30 ± 3.31	0.0 ± 6.16	4.91 ± 3.71	−0.0 ± 5.25	4.07 ± 3.30	0.0 ± 4.90	**3.83 ± 3.05**

For baseline results, the predicted age was generally underestimated for the unseen UKB dataset, with ME as low as −18.20 years for Model 2 on RIG+GS dataset, and only −0.06 years for Model 4 with Fs+FSL preprocessing. This may be expected, considering the smaller age span of the UKB population compared to the multi-site train set population (cf. Supplementary Table 4). Large error values, up to 40 years, were observed for Model 2. The model exhibited a large bias and errors across all four image preprocessing procedures. Additionally, the variance of predictions was larger for 3D models, as seen from Supplementary Fig. 10.

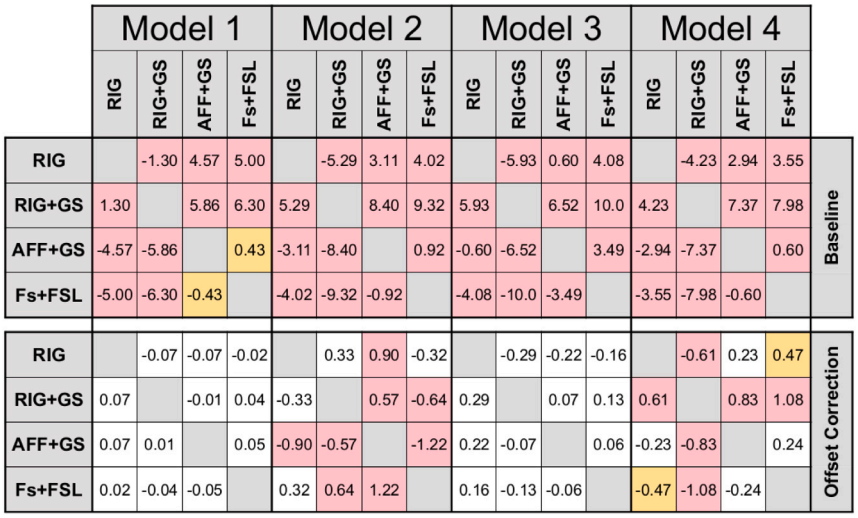
Fig. 6 illustrates the pairwise differences in marginal means of preprocessing, conditional on the model architecture, and the presence or absence of offset correction. For the baseline results, large and significant differences in performance were observed between all combinations of the preprocessing pipelines and models. The disparity in performance is most pronounced between models trained on datasets with affine registration and those with rigid registration; for the former, the MAE nearly doubled. For instance, Model 1, when trained on the RIG+GS dataset, yielded a MAE of 10.52 years, whereas it was 4.22 years with the Fs+FSL preprocessing. For all models, the best baseline result was achieved for Fs+FSL, which can be attributed to the fact that the same software was used for preprocessing of the UKB dataset (Section 2.2).

Correction of systematic offset improved MAE for 4.56 years on average. Despite offset correction, the introduction of different preprocessing of the test set, led to the average increase in MAE 0.5 years, when compared to the results in Table 2 from Section 3.2. This increase was smallest for Model 3 and largest for Model 4. Additionally, the standard deviation of the absolute error increased by up to 0.85 years, and the standard deviation of error also grew by as much as 1.5 years.

In the offset-corrected results, Model 1 generally exhibited best performance, only being surpassed by Model 3 with the RIG preprocessing, achieving an overall best MAE of 3.59 years. For these two architectures, there were no significant differences in performance based on the preprocessing procedure used on the training data (cf. Fig. 6). Despite the differences in datasets and associated preprocessing between the training and testing phases, Model 1 showcased high robustness (MAE variation of 0.07 years), while Model 4 displayed optimal performance prior to offset correction and only matched Model 1 on the Fs+FSL dataset after applying the offset correction. Hence, Model 4 seems susceptible to differences due to preprocessing. As previously, Model 2 was outperformed by the three 3D models. Nevertheless, the introduction of affine registration enhanced the Model 2's performance.

4. Discussion

This work studied the effect of four different T1w preprocessing procedures and implementations on the brain age prediction accuracy using deep learning-based models. For this purpose we implemented, trained and evaluated four CNN architectures presented in the brain age literature.



Baseline	Model 1 RIG	Model 1 RIG+GS	Model 1 AFF+GS	Model 1 Fs+FSL	Model 2 RIG	Model 2 RIG+GS	Model 2 AFF+GS	Model 2 Fs+FSL	Model 3 RIG	Model 3 RIG+GS	Model 3 AFF+GS	Model 3 Fs+FSL	Model 4 RIG	Model 4 RIG+GS	Model 4 AFF+GS	Model 4 Fs+FSL
RIG		-1.30	4.57	5.00		-5.29	3.11	4.02		-5.93	0.60	4.08		-4.23	2.94	3.55
RIG+GS	1.30		5.86	6.30	5.29		8.40	9.32	5.93		6.52	10.0	4.23		7.37	7.98
AFF+GS	-4.57	-5.86		0.43	-3.11	-8.40		0.92	-0.60	-6.52		3.49	-2.94	-7.37		0.60
Fs+FSL	-5.00	-6.30	-0.43		-4.02	-9.32	-0.92		-4.08	-10.0	-3.49		-3.55	-7.98	-0.60	

Offset Correction	Model 1 RIG	Model 1 RIG+GS	Model 1 AFF+GS	Model 1 Fs+FSL	Model 2 RIG	Model 2 RIG+GS	Model 2 AFF+GS	Model 2 Fs+FSL	Model 3 RIG	Model 3 RIG+GS	Model 3 AFF+GS	Model 3 Fs+FSL	Model 4 RIG	Model 4 RIG+GS	Model 4 AFF+GS	Model 4 Fs+FSL
RIG		-0.07	-0.07	-0.02		0.33	0.90	-0.32		-0.29	-0.22	-0.16		-0.61	0.23	0.47
RIG+GS	0.07		-0.01	0.04	-0.33		0.57	-0.64	0.29		0.07	0.13	0.61		0.83	1.08
AFF+GS	0.07	0.01		0.05	-0.90	-0.57		-1.22	0.22	-0.07		0.06	-0.23	-0.83		0.24
Fs+FSL	0.02	-0.04	-0.05		0.32	0.64	1.22		0.16	-0.13	-0.06		-0.47	-1.08	-0.24	

Fig. 6. The pairwise difference in marginal means between preprocessing procedures conditional on model architecture in the presence (*lower*) or absence (*upper*) of offset correction on the unseen site dataset. The preprocessing procedure of the dataset **differed** from preprocessing procedure of the multi-site training dataset. The color of each square marks the significance of difference: red for $p < 0.001$, orange for $0.01 \leq p < 0.001$, yellow for $0.05 \leq p < 0.01$ and white for $p > 0.05$ (not significant).

4.1. Impact of T1w image preprocessing and model architecture

In comparing the four T1w preprocessing pipelines, the most complex AFF+GS pipeline consistently yielded slightly higher brain age prediction accuracy across all models, though this difference was not significant. Increasing the complexity of registration was shown as beneficial in study by Peng et al. [18], wherein the T1w preprocessing procedures including either linear or non-linear registration were compared, resulting in slight favor of the latter. Conversely, Dartora et al. [21] found that a ResNet-based model trained on minimally preprocessed MRIs outperformed models trained with extensive preprocessing. We hypothesize that this superiority may be due to the effects of skull-stripping rather than the complexity of registration and inclusion of gray scale corrections. Skull-stripping, by removing surrounding tissues, potentially eliminates contextual information related to cerebrospinal fluid levels, which are known to increase with age [44,45].

Notably, the inclusion of gray scale correction into the pipeline, i.e. denoising and intensity inhomogeneity correction, while not directly improving MAE, was needed for accurate (linear) image registration used in our preprocessing pipelines. On the other hand good spatial normalization of the input T1w scans eliminates the inter-subject variance due to head size differences and MRI-acquisition related geometric artifacts. As such it particularly benefited Model 2 and 3, receiving downsampled input information.

Discussing computational complexity, Tanveer et al. [46] advocate the shift towards 2D CNN brain age prediction models for use on routine MRIs, with minimal preprocessing. However, our findings contrast this perspective, showing a significant inferiority of the implemented 2D model versus all tested 3D models. The 2D model achieved comparable performance only when trained with the AFF+GS preprocessed dataset. This finding is in line with Feng et al. [20], who showed that a 2D model, which is designed analogous to a 3D model, performs significantly worse. Therefore any future 2D implementations cannot be naive re-implementations of the 3D models, but need to introduce a methodological improvement, like for instance Jönemo et al. [47], predicting age from 2D projections of the 3D MRI volumes.

Preprocessing is generally more computationally demanding than the brain age inference using DL model. The implemented preprocessing pipelines took from 1.5 (RIG) to 16.5 min (AFF+GS), which seems reasonable even for practical implementations. In contrast, model inference for brain age prediction takes only a few seconds, which is negligible compared to the preprocessing times. However, in resource-limited situations (e.g. no GPU), there is a trade-off between implementability and accuracy, as noted by Dartora et al. [21]. Therefore, when selecting the optimal model-pipeline combination, it could become crucial to balance computational requirements and resources against the desired level of accuracy.

Models trained on the Fs+FSL preprocessing, using common software such as FSL and FreeSurfer, presented higher MAE scores than those trained using the AFF+GS pipeline, despite both methods incorporating affine registration and grayscale corrections. Although the differences were not significant, this observation implies that the choice of software could still have a substantial impact on the observed performances, as also highlighted in neuroimaging cortical surface analyses by Kharabian Masouleh et al. [30] and further supported by Bhagwat et al. [31]. It also implies that the results from brain age prediction models, which are trained on the same data but with different T1w preprocessing software implementations, might not be directly comparable.

Comparing the performances reported in the original papers where the four tested CNN models were introduced is challenging, not to mention the variations in training dataset sizes and age structures. For instance, the MAE of Model 1 reported herein was 1 year lower than the MAE reported by Cole et al. [22], even though both used similar T1w preprocessing (RIG), and had comparable training set structures and sizes. We attribute the improvement partially to the mean ensembling and largely to extensive hyperparameter tuning, and the implementation of the preprocessing.

Reproducibility of DL model predictions is critically governed by the availability of the datasets and method implementations, an observation that is supported by the results of this study. For this purpose we applied public datasets and provided the lists of included subject IDs and the exact dataset split assignments as used in this study, as well as the implementations and dependencies of the T1w preprocessing routines, brain age models, scripts to re-run the experiments and carry out the performance evaluations and statistical analyses, all disclosed at the public GitHub repository https://github.com/AralRalud/BrainAgePreprocessing.

4.2. Performance on unseen site dataset with possible new preprocessing

New scanner data, unseen during model training, will generally introduce bias into the brain age estimates of DL models. The increase in MAE on the unseen dataset varied from 0.11 to 1.87 years in comparison to multi-site dataset, varying based on the model and

chosen preprocessing. In related research, Feng et al. [20] reported a rather small increase in MAE of 0.15 years. Multiple other deep learning studies indicate that this increase (or accuracy deterioration) to be much larger. Jonsson et al. [33] reported an increase in MAE of about 3 and 5 years on two separate unseen site datasets. Drop in brain age prediction accuracy was reported also for models trained on datasets with minimal T1w preprocessing. Dartora et al. [21] reported a 1 and 3 year increase in MAE on two independent datasets. Fisch et al. [27] report a 5 year increase on three datasets, prior to applying transfer learning.

While models can partly mitigate differences in preprocessing between training and test sets, both bias and variance will increase. Introducing new preprocessing, different from that used in training, resulted in MAE values with RIG and/or RIG+GS pipelines more than doubling compared to those obtained with Fs+FSL and/or AFF+GS pipelines, with the increase in MAE ranging from 13.91 to 0.52 years. Focusing solely on full-resolution 3D Models 1 and 4, trained on AFF+GS and Fs+FSL pipelines, this increase ranged from only 1.69 to 0.52 years, even before offset correction. These best results align with previously mentioned values, though none of the related studies considered the increase in difference due to changes in preprocessing. We attribute this to the similarity of the T1w preprocessing pipeline (and software) applied to the UKB dataset, as well as generally observed better performance of the models trained with the Fs+FSL and/or AFF+GS pipelines.

However, the increase in MAE seems intrinsically connected to the previously unseen dataset and/or new (unseen) preprocessing procedure, causing domain shift. Domain shift induced deviations of the DL model based age prediction from the actual age may involve a large systematic bias that can be corrected with a simple mitigation strategy. For instance, our model predictions were corrected for age-independent constant offset. Though bias correction, t. i. fitting a linear regression to predictions on validation or test sets, is commonly used in the literature [18,26,29,48–50], several recent studies have cautioned against it [51,52]. Unlike fitting a linear regression line, offset correction does not correct for model's inability to capture linear trend, nor reduces prediction dispersion.

In contrast to the differences in MAE predictions on multi site-dataset, where only marginally significant differences were observed, the differences between models and preprocessing procedures were indeed clearly significant when inferring on the unseen dataset. Among the T1w preprocessing procedures evaluated, those with higher complexity of registration (Fs+FSL and AFF+GS) exhibited the lowest brain age prediction errors when predicting on unseen site dataset, both before and after offset correction. These results are in line with the observation by Cole et al. [22], who found a substantially reduced between-scanner reliability for a model trained on minimally preprocessed T1w images, although Model 1 displayed equivalent performance across all preprocessing pipelines. This shows that, the extensive preprocessing can improve accuracy and acts as a form of data harmonization for unseen site datasets, ensuring consistent predictions.

It is worth noting that the size and age structure in the dataset may adversely influence the brain age prediction accuracy. For example, before offset correction, the models tended to underestimate the age of the subjects in the UKB dataset, which could be due to the younger age of the individuals in the training dataset. Additionally, the observed MAE values on the multi-site and unseen site dataset after offset correction were comparable, which may be partially attributed to the smaller age range of the subjects in the UKB dataset. However, MAE is unlikely to increase proportionally with the widening of the age range in adult datasets, as assumed by Cole et al. [53]. For instance, in an experiment conducted by Peng et al. [18], Model 4, when trained on the UKB and in a separate experiment on a dataset with ages ranging from 17 to 90 years of somewhat similar sizes (i.e., 2600 and 2200 subjects), achieved MAE values of 2.76 and 2.9 years, respectively.

5. Conclusion

In this paper we studied the effect of preprocessing procedure of T1w MRIs on the prediction accuracy of deep brain age models. We considered four preprocessing pipelines, which differed in the degree of freedom of T1w to brain atlas registration, the level of gray scale corrections and software implementations used. Our results for four different CNN architecture show that the choice of software implementation resulted in significant increase in MAE, up to 0.75 years for the same model and dataset. We further show that applying the grayscale corrections does not significantly improve MAE of model predictions. The type of registration was shown to significantly improve MAE when using affine compared to the rigid registration. Models trained on images with isotropic $1 \times 1 \times 1$ mm^3 spacing were less sensitive to the type of T1w preprocessing than the 2D model or the model trained on downsampled 3D images. Most affected by the (mis)registration of the input T1w MRI was the 2D model, since it was limited to only 15 axial slices, predefined in the MNI brain atlas space. In this case, the affine registration was crucial for good model performance, especially when predicting brain age on new dataset not seen during model training. Despite assumptions that models trained on less processed data are better suited for brain age prediction on new scanner datasets, not seen in model training, our results show that extensive T1w preprocessing in fact improves the generalization of brain age models when applied on an unseen datasets. Regardless of the model or the T1w preprocessing used, offset correction should be applied whenever predicting brain age on a new dataset with either the same or different T1w preprocessing as the one used in model training.

CRediT authorship contribution statement

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

Declaration of competing interest

None Declared

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

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

Acknowledgments

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

- **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.

- **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).
- **Cambridge Centre for aging 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 Longitudinal.** Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.
- **the UK Biobank Resource** under Application Number 68981.

Appendix A. Dataset and model details

A.1. Dataset details

See Table 4 and Fig. 7.

A.2. Hyperparameter tuning and selection of loss function

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

Supplementary Fig. 8 presents median, minimal and maximal MAE values of the last 10 epochs for each hyperparameter setting. By choosing the model with smallest median MAE in the last 10 epochs we could identify hyperparameter setting, with which the training converged well. Due to GPU space constraints, the maximal batch size was 24 for Model 1 and 9 for Model 4.

For regression Model 1, 2 and 3, training with the MSE loss often diverged for larger learning rate values; this was also the case for Models 2 and 3 with the learning rate values set as proposed in original papers. In general, we observed that training with L1 loss was most stable and produced overall lower MAE values, compared to the use of Mean-Squared Error and Kullback–Leibler divergence losses. Hence, hereafter we used the L1 loss in regression Models 1, 2, 3. The chosen optimal hyperparameter values and the original and resulting model accuracy are given in Supplementary Table 5.

Unless noted otherwise, we used the hyperparameters reported in Supplementary Table 5. in all subsequent experiments. Models based on these hyperparameters represent our baseline models.

Table 4
Age statistics, i.e. span, mean age (μ_{age}), and associated standard deviation ($\mathbf{sd}_{age}$) in *years*, per dataset included in the train, test, and validation datasets (*top*), and the unseen site datasets (*bottom*).

Aim: Train, validation, test (Multi-site T1w scans)			
Dataset	$\mathbf{N}_{scans}$	Age span	$\mu_{age} \pm \mathbf{sd}_{age}$
ABIDE I[a]	161	18.0–48.0	25.7 ± 6.4
ADNI[b,h]	248	60.0–90.0	76.2 ± 5.1
CamCAN [54,55][c]	624	18.0–88.0	54.2 ± 18.4
CC-359 [56][d]	349	29.0–80.0	53.5 ± 7.8
FCON 1000[e]	572	18.0–85.0	45.3 ± 18.9
IXI[f]	472	20.1–86.2	49.0 ± 16.2
OASIS-2 [57][g]	78	60.0 – 95.0	75.6 ± 8.4
Total	2504	18.0–95.0	52.1 ± 19.1
Aim: Test (Unseen T1w scans)			
Dataset	$\mathbf{N}_{subj}$	Age span	$\mu_{age} \pm \mathbf{sd}_{age}$
UK Biobank [58]	1493	48.5–80.4	63.1 ± 7.2

[a] Data available at: http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html.
[b] Data available at: http://adni.loni.usc.edu/.
[c] Data available at: https://camcan-archive.mrc-cbu.cam.ac.uk/dataaccess/.
[d] Data available at: https://sites.google.com/view/calgary-campinas-dataset/download.
[e] Data available at: http://fcon_1000.projects.nitrc.org/indi/enhanced/neurodata.html.
[f] Data available at: https://brain-development.org/ixi-dataset/.
[g] Data available at: https://www.oasis-brains.org/.
[h] 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).

A.3. Model predictions on unseen site dataset

The Supplementary Fig. 9 and Supplementary Fig. 10 show model predictions on UKB dataset using the same (Section 3.2) and different preprocessing (Section 3.3) as used on the training set. The predictions show a clear systematic offset, specific to each combination of preprocessing and model architecture.

A.4. Execution times

All experiments were run on the same workstation with Intel Core i7-8700K CPU, 64 GB system memory and three NVIDIA GeForce RTX 2080 Ti GPUs, each with 11 GB dedicated memory. The image preprocessing pipelines and model architectures differed based on their execution and training time, respectively, and the hardware requirements (cf. Table 6). The RIG preprocessing pipeline took < 2 min, while the more complex AFF+GS took 4-7 min per image. The Fs+FSL pipeline was most time consuming, taking > 15 min per image on average.

The difference in both the model training time and the hardware requirements is substantial for different model architectures. Models 1 and 4, trained on full resolution input 3D images require more than twice as much training time and GPU memory, compared to Models 2 and 4. Despite the larger number of trainable parameters in Model 3, its accuracy and robustness were comparable to that of Models 1 and 4.

Appendix B. Linear mixed effect model results

The subsequent section presents detailed results from the LMEM and ANOVA tests corresponding to specific experiments. Specifically, refer to Table 7 for Section 3.1, Table 8 for Section 3.2, and Table 9 for Section 3.3. The levels of statistical significance are denoted as: '$***$' for $0 < p < 0.001$, '$**$' for $0.001 < p < 0.01$ and '$*$' for $0.01 < p < 0.05$.

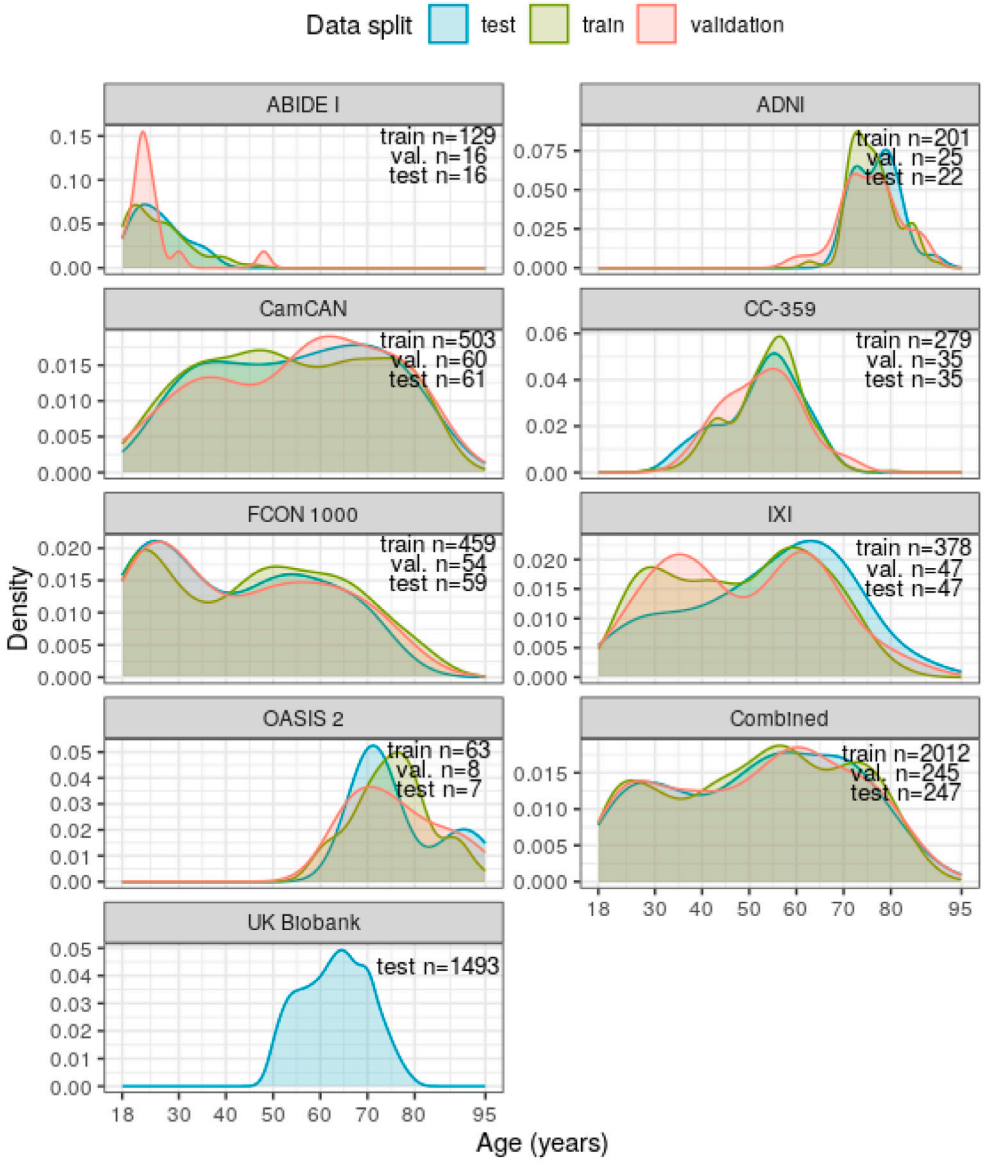


Fig. 7. Density of age distribution per each dataset and combined multi-site dataset, depicted for train, test and validation set splits.

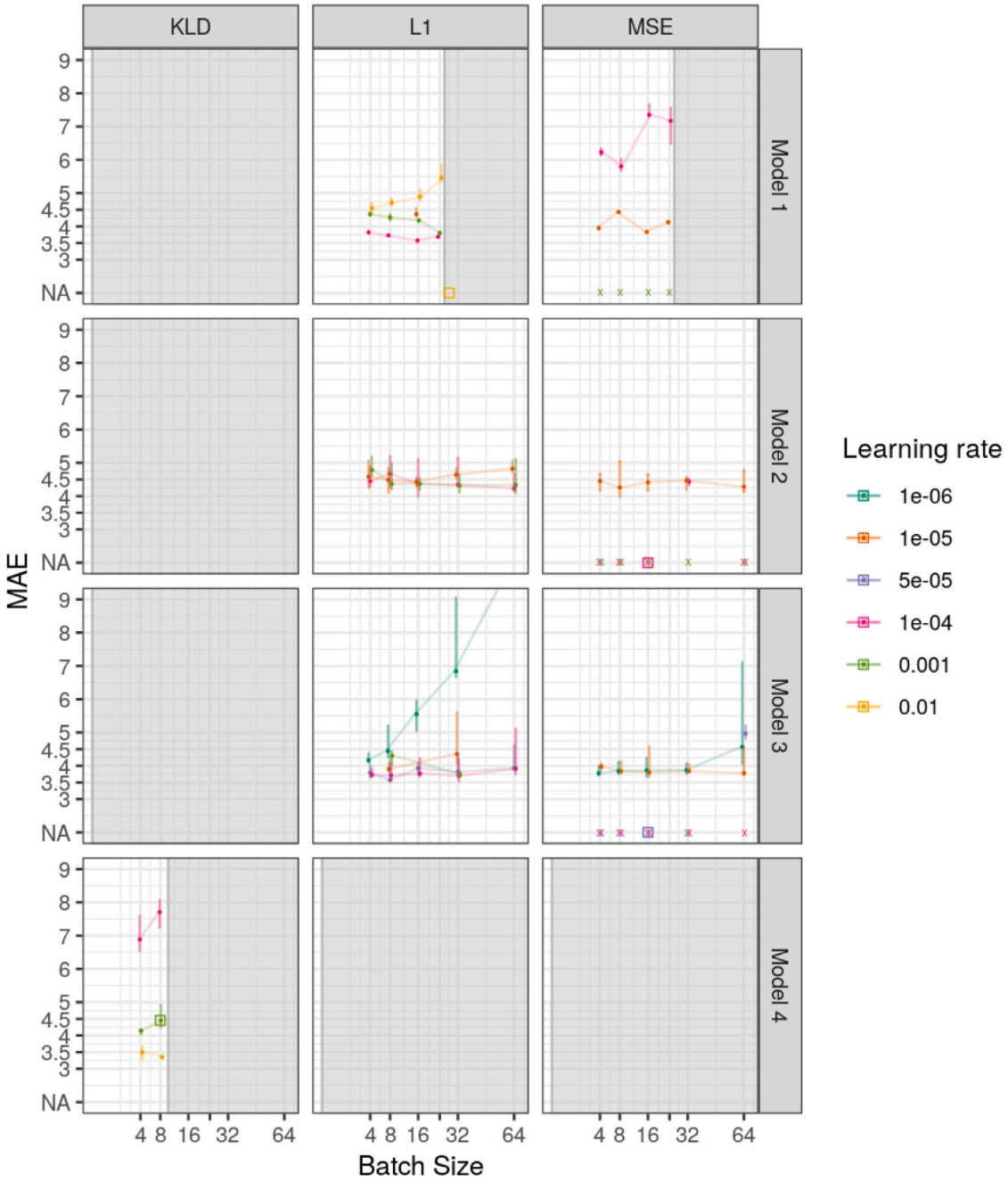


Fig. 8. Median, minimal and maximal MAE value of 10 last training epochs for each hyperparameter setting. The hyperparameter values proposed in original research of four models are marked with *square*, the ones resulting in training divergence are *marked as NA and with a cross*. Hyperparameter space for large batch sizes was inaccessible due to hardware limitations and is *grayed out*.

Table 5
Proposed hyperparameter values in original literature and the values chosen herein. The resulting model accuracy is reported as MAE in years.

	Model 1		Model 2	
	Proposed	Implemented	Proposed	Implemented
Input size	$182 \times 218 \times 182$	$157 \times 189 \times 170$	$157 \times 189 \times 15$	
[a]Batch size	28	16	16	32
[a]Loss function	L1		MSE	L1
[a]Learning rate	1×10^{-2}	1×10^{-4}	1×10^{-4}	1×10^{-3}
Learning rate 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]$med[min, max]$	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	$95 \times 79 \times 78$		$160 \times 192 \times 160$	$157 \times 189 \times 170$
[a]Batch size	16	8	8	8
[a]Loss function	MSE	L1	Kullback–Leibler divergence	
[a]Learning rate	5×10^{-5}		1×10^{-2}	
Learning rate 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]$med[min, max]$	3.67	3.57 [3.52, 4.26]	2.14	3.35 [3.29, 3.42]

[a] Hyperparameters were reevaluated.

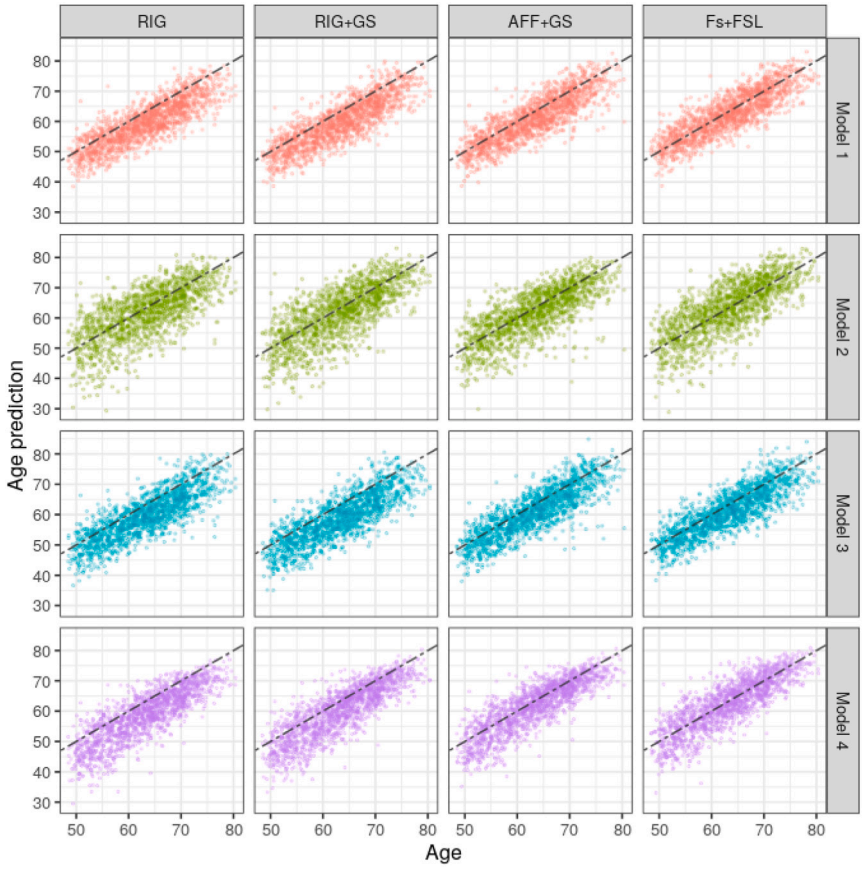


Fig. 9. Model predictions on UKB dataset, preprocessed in the same manner as the training set.

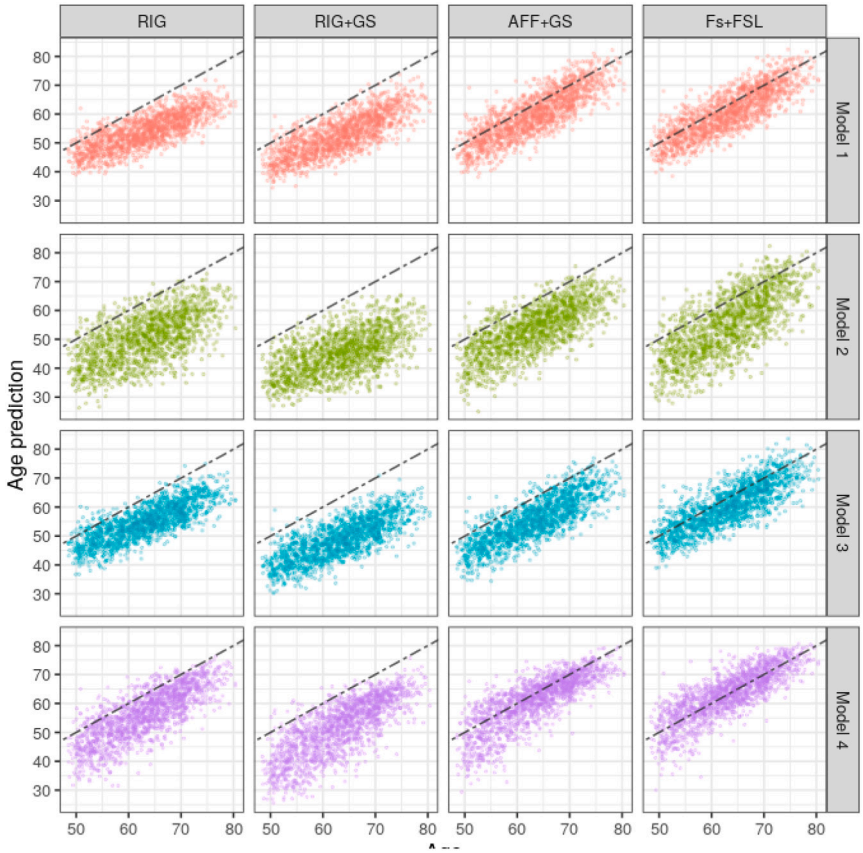


Fig. 10. Model predictions on UKB dataset, preprocessed using different preprocessing pipeline as the training set.

Table 6
Average run time of preprocessing pipeline per image (*left*) and model training times with hardware requirements (*right*).

Image preprocessing	Time [m:ss]	Model	Time [h]	No. GPUs
RIG	1:25	Model 1	15.5	2
RIG+GS	6:30	Model 2	8.9	1
AFF+GS	7:40	Model 3	7.17	1
Fs+FSL	16:20	Model 4	20.2	3

Table 7
Results of ANOVA and LMEM with absolute error as response variable, model architecture and preprocessing procedure as fixed factor on test set of Multi-site dataset: $Abs\ Error = Model + Preprocessing + Model * Preprocessing + (1|ID)$. Interaction was not statistically significant. Here, 'NumDF' denotes the numerator degrees of freedom, and 'DenDF' denotes the denominator degrees of freedom.

ANOVA	NumDF	DenDF	F value	*p*-value	LMEM	Estimate	Std. Err.	2.5%	97.5%
					Intercept	3.181	0.190	2.809	3.553
Model	3	3690	49.506	***	Model 2	1.290	0.181	0.936	1.645
					Model 3	0.265	0.181	−0.090	0.612
					Model 4	0.135	0.181	−0.220	0.489
Preproc.	3	3690	5.090	**	RIG+GS	−0.062	0.181	−0.416	0.293
					AFF+GS	−0.221	0.181	−0.575	0.134
					Fs+FSL	0.036	0.181	−0.319	0.390
Model: Preproc	9	3690	1.143		Model 2:RIG+GS	−0.117	0.256	−0.619	0.384
					Model 3:RIG+GS	0.111	0.256	−0.390	0.612
					Model 4:RIG+GS	0.033	0.256	−0.468	0.534
					Model 2:AFF+GS	−0.533	0.256	−1.034	−0.032
					Model 3:AFF+GS	−0.048	0.256	−0.549	0.454
					Model 4:AFF+GS	0.155	0.256	−0.347	0.656
					Model 2:Fs+FSL	−0.426	0.256	−0.927	0.075
					Model 3:Fs+FSL	−0.033	0.256	−0.535	0.468
					Model 4:Fs+FSL	−0.037	0.256	−0.539	0.464
					Random effects		Variance	SD	
					Subject ID (Intercept)		4.843	2.201	
					Residual		4.055	2.014	

Table 8
Results of ANOVA and LMEM tests on the UK Biobank dataset preprocessed with the same preprocessing procedure as the training dataset with absolute error as response variable, and model architecture, offset correction (OC) and preprocessing procedure as fixed factor: $Abs\ Error = Model + Preproc. + OC + Model * Preproc + Model * OC + Preproc * OC + Model * Preproc * OC + (1|ID)$. Here, 'NumDF' denotes the numerator degrees of freedom, and 'DenDF' denotes the denominator degrees of freedom.

ANOVA	NumDF	DenDF	F value	*p*-value	LMEM	Estimate	Std. Err.	2.5%	97.5%
					Intercept	4.478	0.082	4.317	4.640
OC	1	46 252	637.861	***	OC	−1.217	0.091	−1.395	−1.039
Model	3	46 252	356.449	***	Model 2	0.588	0.091	0.410	0.767
					Model 3	0.250	0.091	0.072	0.428
					Model 4	0.712	0.091	0.534	0.890
Preproc.	3	46 252	221.364	***	RIG+GS	−0.010	0.091	−0.188	0.169
					AFF+GS	−0.752	0.091	−0.930	−0.574
					Fs+FSL	−1.152	0.091	−1.330	−0.974
OC: Model	3	46 252	53.303	***	OC:Model 2	1.096	0.129	0.844	1.348
					OC:Model 3	0.095	0.129	−0.157	0.347
					OC:Model 4	−0.280	0.129	−0.532	−0.028
OC: Preproc.	3	46 252	93.765	***	OC:RIG+GS	0.077	0.129	−0.175	0.329
					OC:AFF+GS	0.804	0.129	0.552	1.056
					OC:Fs+FSL	1.171	0.129	0.919	1.423
Model: Preproc.	9	46 252	12.161	***	Model 2:RIG+GS	−0.153	0.129	−0.405	0.099
					Model 3:RIG+GS	0.543	0.129	0.291	0.795
					Model 4:RIG+GS	−0.687	0.129	−0.939	−0.435
					Model 2:AFF+GS	0.010	0.129	−0.242	0.262
					Model 3:AFF+GS	−0.047	0.129	−0.299	0.205
					Model 4:AFF+GS	−0.789	0.129	−1.041	−0.537
					Model 2:Fs+FSL	0.628	0.129	0.376	0.880
					Model 3:Fs+FSL	0.207	0.129	−0.045	0.459
					Model 4:Fs+FSL	−0.410	0.129	−0.662	−0.158
OC: Model: Preproc.	9	46 252	12.981	***	OC:Model 2:RIG+GS	−0.113	0.182	−0.470	0.243
					OC:Model 3:RIG+GS	−0.513	0.182	−0.869	−0.156
					OC:Model 4:RIG+GS	0.634	0.182	0.277	0.990
					OC:Model 2:AFF+GS	−0.796	0.182	−1.152	−0.439
					OC:Model 3:AFF+GS	−0.096	0.182	−0.453	0.260
					OC:Model 4:AFF+GS	0.491	0.182	0.134	0.847
					OC:Model 2:Fs+FSL	−1.052	0.182	−1.408	−0.695
					OC:Model 3:Fs+FSL	−0.410	0.182	−0.766	−0.053
					OC:Model 4:Fs+FSL	0.285	0.182	−0.072	0.641
					Random effects		Variance	SD	
					Subject ID (Intercept)		3.937	1.984	
					Residual		6.174	2.485	

Table 9
Results of ANOVA and LMEM tests on UK Biobank dataset preprocessed with new preprocessing procedure with absolute error as response variable, and model architecture, offset correction (OC) and preprocessing procedure as fixed factor: $Abs\ Error = OC + Model + Preprocessing + OC * Model + OC * Preprocessing + Model * Preprocessing + Model * Preprocessing * OC + (1|ID)$. Here, 'NumDF' denotes the numerator degrees of freedom, and 'DenDF' denotes the denominator degrees of freedom.

ANOVA	NumDF	DenDF	F value	*p*-value	LMEM	Estimate	Std. Err.	2.5%	97.5%
					Intercept	9.228	0.107	9.019	9.436
OC	1	46 252	20 035.70	***	OC	−5.588	0.129	−5.840	−5.335
Model	3	46 252	2514.49	***	Model 2	3.681	0.129	3.428	3.933
					Model 3	−0.694	0.129	−0.947	−0.442
					Model 4	−1.852	0.129	−2.104	−1.599
Preproc	3	46 252	3513.01	***	RIG+GS	1.295	0.129	1.043	1.548
					AFF+GS	−4.574	0.129	−4.826	−4.321
					.Fs+FSL	−5.003	0.129	−5.256	−4.751
OC: Model	3	46 252	946.12	***	OC:Model 2	−1.737	0.182	−2.094	−1.380
					OC:Model 3	0.648	0.182	0.291	1.005
					OC:Model 4	2.513	0.182	2.156	2.870
OC: Preproc	3	46 252	3126.95	***	OC:RIG+GS	−1.227	0.182	−1.585	−0.870
					OC:AFF+GS	4.648	0.182	4.291	5.005
					OC:Fs+FSL	5.027	0.182	4.670	5.384
Model: Preproc	9	46 252	77.89	***	Model 2:RIG+GS	3.996	0.182	3.639	4.354
					Model 3:RIG+GS	4.631	0.182	4.274	4.988
					Model 4:RIG+GS	3.134	0.182	2.777	3.491
					Model 2:AFF+GS	1.466	0.182	1.109	1.823
					Model 3:AFF+GS	3.978	0.182	3.621	4.335
					Model 4:AFF+GS	1.632	0.182	1.275	1.989
					Model 2:Fs+FSL	0.979	0.182	0.622	1.336
					Model 3:Fs+FSL	0.920	0.182	0.563	1.277
					Model 4:Fs+FSL	1.456	0.182	1.099	1.813
OC: Model: Preproc	9	46 252	71.25	***	OC:Model 2:RIG+GS	−4.394	0.258	−4.899	−3.889
					OC:Model 3:RIG+GS	−4.410	0.258	−4.915	−3.905
					OC:Model 4:RIG+GS	−2.594	0.258	−3.099	−2.090
					OC:Model 2:AFF+GS	−2.443	0.258	−2.948	−1.938
					OC:Model 3:AFF+GS	−3.835	0.258	−4.340	−3.330
					OC:Model 4:AFF+GS	−1.933	0.258	−2.438	−1.428
					OC:Model 2:Fs+FSL	−0.686	0.258	−1.191	−0.181
					OC:Model 3:Fs+FSL	−0.787	0.258	−1.292	−0.282
					OC:Model 4:Fs+FSL	−1.951	0.258	−2.456	−1.446
					Random effects		Variance	SD	
					Subject ID (Intercept)		4.577	2.139	
					Residual		12.396	3.521	

References

[1] K. Franke, C. Gaser, Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? Front. Neurol. 10 (2019) 789.

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

[3] E.A. Høgestøl, T. Kaufmann, G.O. Nygaard, M.K. Beyer, P. Sowa, J.E. Nordvik, K. Kolskår, G. Richard, O.A. Andreassen, H.F. Harbo, L.T. Westlye, Cross-sectional and longitudinal MRI brain scans reveal accelerated brain aging in multiple sclerosis, Front. Neurol. 10 (2019).

[4] J.H. Cole, J. Raffel, T. Friede, A. Eshaghi, W.J. Brownlee, D. Chard, N.D. Stefano, C. Enzinger, L. Pirpamer, M. Filippi, C. Gasperini, M.A. Rocca, A. Rovira, S. Ruggieri, J. Sastre-Garriga, M.L. Stromillo, B.M.J. Uitdehaag, H. Vrenken, F. Barkhof, R. Nicholas, O. Ciccarelli, Longitudinal assessment of multiple sclerosis with the brain-age paradigm, Ann. Neurol. 88 (1) (2020) 93–105.

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

[6] K.J. Petersen, N. Metcalf, S. Cooley, D. Tomov, F. Vaida, R. Paul, B.M. Ances, Accelerated brain aging and cerebral blood flow reduction in persons with human immunodeficiency virus, Clin. Infect. Dis. 73 (10) (2021) 1813–1821.

[7] J.H. Cole, J. Underwood, M.W.A. Caan, D.D. Francesco, R.A.v. Zoest, R. Leech, F.W.N.M. Wit, P. Portegies, G.J. Geurtsen, B.A. Schmand, M.F.S.v.d. Loeff, C. Franceschi, C.A. Sabin, C.B.L.M. Majoie, A. Winston, P. Reiss, D.J. Sharp, Increased brain-predicted aging in treated HIV disease, Neurology 88 (14) (2017) 1349–1357.

[8] K. Ning, L. Zhao, W. Matloff, F. Sun, A.W. Toga, Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants, Sc. Rep. 10 (1) (2020) 10.

[9] Z. Linli, J. Feng, W. Zhao, S. Guo, Associations between smoking and accelerated brain ageing, Prog. Neuro-Psychopharmacol. Biol. Psych. 113 (2022) 110471.

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

[11] F. Liem, G. Varoquaux, J. Kynast, F. Beyer, S. Kharabian Masouleh, J.M. Huntenburg, L. Lampe, M. Rahim, A. Abraham, R.C. Craddock, S. Riedel-Heller, T. Luck, M. Loeffler, M.L. Schroeter, A.V. Witte, A. Villringer, D.S. Margulies, Predicting brain-age from multimodal imaging data captures cognitive impairment, NeuroImage 148 (2017) 179–188.

[12] I. Hwang, E.K. Yeon, J.Y. Lee, R.-E. Yoo, K.M. Kang, T.J. Yun, S.H. Choi, C.-H. Sohn, H. Kim, J.-h. Kim, Prediction of brain age from routine T2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network, Neurobiol. Aging 105 (2021) 78–85.

[13] G. Richard, K. Kolskår, A.-M. Sanders, T. Kaufmann, A. Petersen, N.T. Doan, J.M. Sánchez, D. Alnæs, K.M. Ulrichsen, E.S. Dørum, O.A. Andreassen, J.E. Nordvik, L.T. Westlye, Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry, PeerJ 6 (2018) e5908.

[14] S. Tønnesen, T. Kaufmann, A.-M.G. de Lange, G. Richard, N.T. Doan, D. Alnæs, D. van der Meer, J. Rokicki, T. Moberget, I.I. Maximov, I. Agartz, S.R. Aminoff, D. Beck, D.M. Barch, J. Beresniewicz, S. Cervenka, H. Fatouros-Bergman, A.R. Craven, L. Flyckt, T.P. Gurholt, U.K. Haukvik, K. Hugdahl, E. Johnsen, E.G. Jönsson, L. Farde, L. Flyckt, G. Engberg, S. Erhardt, H. Fatouros-Bergman, S. Cervenka, L. Schwieler, F. Piehl, I. Agartz, K. Collste, P. Victorsson, A. Malmqvist, M. Hedberg, F. Orhan, C. Sellgren, K.K. Kolskår, R.A. Kroken, T.V. Lagerberg, E.-M. Løberg, J.E. Nordvik, A.-M. Sanders, K. Ulrichsen, O.A. Andreassen, L.T. Westlye, Brain age prediction reveals aberrant brain white matter in schizophrenia and bipolar disorder: A multisample diffusion tensor imaging study, Biol. Psychiatry: Cogn. Neurosci. Neuroimaging 5 (12) (2020) 1095–1103.

[15] J. Gao, J. Liu, Y. Xu, D. Peng, Z. Wang, Brain age prediction using the graph neural network based on resting-state functional MRI in Alzheimer's disease, Front. Neurosci. 17 (2023) 1222751.

[16] I. Beheshti, S. Nugent, O. Potvin, S. Duchesne, Disappearing metabolic youthfulness in the cognitively impaired female brain, Neurobiol. Aging 101 (2021) 224–229.

[17] P.K. Lam, V. Santhalingam, P. Suresh, R. Baboota, A.H. Zhu, S.I. Thomopoulos, N. Jahanshad, P.M. Thompson, Accurate brain age prediction using recurrent slice-based networks, in: J. Brieva, N. Lepore, M.G. Linguraru, E.R.C. M.D. (Eds.), 16th International Symposium on Medical Information Processing and Analysis, Vol. 11583, International Society for Optics and Photonics, SPIE, 2020, 1158303, http://dx.doi.org/10.1117/12.2579630.

[18] H. Peng, W. Gong, C.F. Beckmann, A. Vedaldi, S.M. Smith, Accurate brain age prediction with lightweight deep neural networks, Med. Image Anal. 68 (2021).

[19] B. Dufumier, P. Gori, I. Battaglia, J. Victor, A. Grigis, E. Duchesnay, Benchmarking CNN on 3D anatomical brain MRI: Architectures, data augmentation and deep ensemble learning, 2021, http://dx.doi.org/10.48550/ARXIV.2106.01132, URL: https://arxiv.org/abs/2106.01132.

[20] X. Feng, Z.C. Lipton, J. Yang, S.A. Small, F.A. Provenzano, Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging, Neurobiol. Aging 91 (2020) 15–25.

[21] C. Dartora, A. Marseglia, G. Mårtensson, G. Rukh, J. Dang, J.-S. Muehlboeck, L.-O. Wahlund, R. Moreno, J. Barroso, D. Ferreira, H.B. Schiöth, E. Westman, for the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing, the Japanese Alzheimer's Disease Neuroimaging Initiative, the AddNeuroMed Consortium, A deep learning model for brain age prediction using minimally preprocessed t1w images as input, Front. Aging Neurosci. 15 (2024).

[22] J.H. Cole, R.P.K. Poudel, D. Tsagkrasoulis, M.W.A. Caan, C. Steves, T.D. Spector, G. Montana, Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker, NeuroImage 163 (2017) 115–124.

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

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

[25] K.-M. Bintsi, V. Baltatzis, A. Kolbeinsson, A. Hammers, D. Rueckert, Patch-based brain age estimation from MR images, 2020, URL: http://arxiv.org/abs/2008.12965.

[26] J. Cheng, Z. Liu, H. Guan, Z. Wu, H. Zhu, J. Jiang, W. Wen, D. Tao, T. Liu, Brain age estimation from MRI using cascade networks with ranking loss, IEEE Trans. Med. Imaging 40 (12) (2021) 3400–3412.

[27] L. Fisch, J. Ernsting, N.R. Winter, V. Holstein, R. Leenings, M. Beisemann, K. Sarink, D. Emden, N. Opel, R. Redlich, J. Repple, D. Grotegerd, S. Meinert, N. Wulms, H. Minnerup, J.G. Hirsch, T. Niendorf, B. Endemann, F. Bamberg, T. Kröncke, A. Peters, R. Bülow, H. Völzke, O. von Stackelberg, R.F. Sowade, L. Umutlu, B. Schmidt, S. Caspers, Consortium, German National Cohort Study Center, H. Kugel, B.T. Baune, T. Kircher, B. Risse, U. Dannlowski, K. Berger, T. Hahn, Predicting brain-age from raw T 1 -weighted magnetic resonance imaging data using 3D convolutional neural networks, 2021, http://dx.doi.org/10.48550/ARXIV.2103.11695, URL: https://arxiv.org/abs/2103.11695.

[28] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, R. Horaud, A comprehensive analysis of deep regression, IEEE Trans. Pattern Anal. Mach. Intell. 42 (9) (2020) 2065–2081.

[29] J.H. Cole, T. Annus, L.R. Wilson, R. Remtulla, Y.T. Hong, T.D. Fryer, J. Acosta-Cabronero, A. Cardenas-Blanco, R. Smith, D.K. Menon, S.H. Zaman, P.J. Nestor, A.J. Holland, Brain-predicted age in Down syndrome is associated with beta amyloid deposition and cognitive decline, Neurobiol. Aging 56 (2017) 41–49.

[30] S. Kharabian Masouleh, S.B. Eickhoff, Y. Zeighami, L.B. Lewis, R. Dahnke, C. Gaser, F. Chouinard-Decorte, C. Lepage, L.H. Scholtens, F. Hoffstaedter, D.C. Glahn, J. Blangero, A.C. Evans, S. Genon, S.L. Valk, Influence of processing pipeline on cortical thickness measurement, Cereb Cortex 30 (9) (2020) 5014–5027.

[31] N. Bhagwat, A. Barry, E.W. Dickie, S.T. Brown, G.A. Devenyi, K. Hatano, E. DuPre, A. Dagher, M. Chakravarty, C.M.T. Greenwood, B. Misic, D.N. Kennedy, J.-B. Poline, Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses, GigaScience 10 (1) (2021).

[32] M. de Fátima Machado Dias, P. Carvalho, M. Castelo-Branco, J. Valente Duarte, Cortical thickness in brain imaging studies using FreeSurfer and CAT12: A matter of reproducibility, Neuroimage: Rep. 2 (4) (2022) 100137.

[33] B.A. Jonsson, G. Bjornsdottir, T.E. Thorgeirsson, L.M. Ellingsen, G.B. Walters, D.F. Gudbjartsson, H. Stefansson, K. Stefansson, M.O. Ulfarsson, Brain age prediction using deep learning uncovers associated sequence variants, Nature Commun. 10 (1) (2019) 5409.

[34] J.V. Manjón, P. Coupé, L. Martí-Bonmatí, D.L. Collins, M. Robles, Adaptive non-local means denoising of MR images with spatially varying noise levels, J. Magn. Reson. Imaging 31 (1) (2010) 192–203.

[35] V. Fonov, A. Evans, R. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage 47 (2009) S102.

[36] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, J.C. Gee, N4ITK: improved N3 bias correction, IEEE Trans. Med. Imaging 29 (6) (2010) 1310–1320.

[37] M. Modat, D.M. Cash, P. Daga, G.P. Winston, J.S. Duncan, S. Ourselin, Global image registration using a symmetric block-matching approach, J. Med. Imaging 1 (2) (2014) 1–6.

[38] M. Jenkinson, C.F. Beckmann, T.E.J. Behrens, M.W. Woolrich, S.M. Smith, FSL, NeuroImage 62 (2) (2012) 782–790.

[39] FreeSurferWiki, FreeSurferWiki: recon-all, 2022, URL: https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all.

[40] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, Neuroimage 17 (2) (2002) 825–841.

[41] S.M. Smith, F. Alfaro-Almagro, K.L. Miller, UK Biobank Brain Imaging Documentation, Welcome Centre for Integrative Neuroimaging and Oxford University, 2020, URL: https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf.

[42] G. Grabner, A.L. Janke, M.M. Budge, D. Smith, J. Pruessner, D.L. Collins, Symmetric atlasing and model based segmentation: An application to the hippocampus in older adults, in: R. Larsen, M. Nielsen, J. Sporring (Eds.), Medical Image Computing and Computer-Assisted Intervention, MICCAI 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 58–66.

[43] M. Jenkinson, S. Smith, A global optimisation method for robust affine registration of brain images, Med. Image Anal. 5 (2) (2001) 143–156.

[44] A. Irimia, Cross-sectional volumes and trajectories of the human brain, gray matter, white matter and cerebrospinal fluid in 9473 typically aging adults, Neuroinform 19 (2) (2021) 347–366.

[45] S. Yamada, T. Otani, S. Ii, H. Kawano, K. Nozaki, S. Wada, M. Oshima, Y. Watanabe, Aging-related volume changes in the brain and cerebrospinal fluid using artificial intelligence-automated segmentation, Eur. Radiol. 33 (10) (2023) 7099–7112.

[46] M. Tanveer, M. Ganaie, I. Beheshti, T. Goel, N. Ahmad, K.-T. Lai, K. Huang, Y.-D. Zhang, J. Del Ser, C.-T. Lin, Deep learning for brain age estimation: A systematic review, Inf. Fusion 96 (2023) 130–143.

[47] J. Jönemo, M.U. Akbar, R. Kämpe, J.P. Hamilton, A. Eklund, Efficient brain age prediction from 3D MRI volumes using 2D projections, Brain Sci. 13 (9) (2023) 1329.

[48] A.-M.G.d. Lange, T. Kaufmann, D.v.d. Meer, L.A. Maglanoc, D. Alnæs, T. Moberget, G. Douaud, O.A. Andreassen, L.T. Westlye, Population-based neuroimaging reveals traces of childbirth in the maternal brain, Proc. Natl. Acad. Sci. USA 116 (44) (2019) 22341–22346.

[49] S.M. Smith, D. Vidaurre, F. Alfaro-Almagro, T.E. Nichols, K.L. Miller, Estimation of brain age delta from brain imaging, NeuroImage 200 (2019) 528–539.

[50] T. Dunås, A. Wåhlin, L. Nyberg, C.-J. Boraxbekk, Multimodal image analysis of apparent brain age identifies physical fitness as predictor of brain maintenance, Cerebral Cortex (bhab019) (2021).

[51] E.R. Butler, A. Chen, R. Ramadan, T.T. Le, K. Ruparel, T.M. Moore, T.D. Satterthwaite, F. Zhang, H. Shou, R.C. Gur, T.E. Nichols, R.T. Shinohara, Pitfalls in brain age analyses, Hum. Brain Map. 42 (13) (2021) 4092–4101.

[52] A.-M.G. de Lange, M. Anatürk, J. Rokicki, L.K.M. Han, K. Franke, D. Alnæ s, K.P. Ebmeier, B. Draganski, T. Kaufmann, L.T. Westlye, T. Hahn, J.H. Cole, Mind the gap: Performance metric evaluation in brain-age prediction, Hum. Brain Map. 43 (10) (2022) 3113–3129.

[53] J.H. Cole, K. Franke, N. Cherbuin, Quantification of the biological age of the brain using neuroimaging, in: A. Moskalev (Ed.), Biomarkers of Human Aging, in: Healthy Ageing and Longevity, Springer International Publishing, Cham, 2019, pp. 293–328, http://dx.doi.org/10.1007/978-3-030-24970-0_19.

[54] M.A. Shafto, L.K. Tyler, M. Dixon, J.R. Taylor, J.B. Rowe, R. Cusack, A.J. Calder, W.D. Marslen-Wilson, J. Duncan, T. Dalgleish, R.N. Henson, C. Brayne, F.E. Matthews, The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing, BMC Neurol. 14 (2014).

[55] J.R. Taylor, N. Williams, R. Cusack, T. Auer, M.A. Shafto, M. Dixon, L.K. Tyler, n. Cam-Can, R.N. Henson, 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) (2017) 262–269.

[56] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, R. Lotufo, An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement, NeuroImage 170 (2018) 482–494.

[57] D.S. Marcus, A.F. Fotenos, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults, J. Cogn. Neurosci. 22 (12) (2010) 2677–2684.

[58] K.L. Miller, F. Alfaro-Almagro, N.K. Bangerter, D.L. Thomas, E. Yacoub, J. Xu, A.J. Bartsch, S. Jbabdi, S.N. Sotiropoulos, J.L.R. Andersson, L. Griffanti, G. Douaud, T.W. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P.M. Matthews, S.M. Smith, Multimodal population brain imaging in the UK Biobank prospective epidemiological study, Nat. Neurosci. 19 (11) (2016) 1523–1536.