Highlights

Extensive T1-weighted MRI Preprocessing Improves Generalizability of Deep Brain Age Prediction Models

Lara Dular,Franjo Pernuš,Žiga Špiclin

- This study involved a thorough and reproducible quantitative assessment of the impact of four T1w preprocessing variants on brain age prediction accuracy using four recent deep learning-based model architectures.
- Repeated model training with random initialization and use of linear mixed-effects models enabled the statistical analysis of the effect of various confounding factors on brain age prediction accuracy.
- The choice of T1w preprocessing software implementation resulted in statistically significant increase in mean absolute error of up to 0.7 years for the same model and dataset.
- Our results show that extensive T1w preprocessing, with higher degree of freedom in T1w to atlas registration and extensive grayscale corrections, and bias correction improve the generalization of brain age models' performances when applied on new unseen datasets.

Extensive T1-weighted MRI Preprocessing Improves Generalizability of Deep Brain Age Prediction Models⋆

Lara Dular[a,*], Franjo Pernuš[a] and Žiga Špiclin[a]

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

ARTICLE INFO

Keywords:
brain age
MRI preprocessing
deep regression model
linear mixed effect models
data bias
transfer learning
reproducible research

Abstract

Brain age is an estimate of chronological age obtained from T1-weighted magnetic resonance images (T1w MRI) and represents a simple diagnostic biomarker of brain ageing and associated diseases. While the current best accuracy of brain age predictions on healthy subject T1w MRIs is from two to three years, comparing results from different studies is challenging due to differences in datasets, T1w preprocessing pipelines, and performance metrics used. This paper investigates the impact of T1w image preprocessing on the performance of four deep learning-based brain age models presented in recent literature. Four preprocessing pipelines were evaluated, differing in terms of registration, grayscale correction, and software implementation. The results showed that the choice of software implementation can significantly affect the prediction error, with a maximum increase of 0.7 years in mean absolute error (MAE) for the same model and dataset. While grayscale correction had no significant impact on MAE, the affine registration, compared to the rigid registration of T1w images to brain atlas was shown to statistically significantly improve MAE. Models trained on 3D images with isotropic 1 mm^3 resolution were less sensitive to T1w preprocessing than 2D models or models trained on downsampled 3D images. Contrary to the indications from research literature that models trained on less preprocessed T1w scans are better suited for age predictions on T1w images from new scanners, not seen in model training, our results show that extensive T1w preprocessing in fact improves the MAE when new dataset is used. Regardless of model or T1w preprocessing used, we show that to enable generalization of model's performance on a new dataset with either the same or different T1w preprocessing than the one applied in model training some form of bias correction should be applied.

1. Introduction

Brain age is obtained by applying regression models to predict chronological age from T1-weighted magnetic resonance images (T1w MRIs). The brain age is expected to equal chronological age in healthy subjects, while in the presence of a certain pathology or chronic disease the difference or gap between the two is likely increased. For instance, premature brain ageing has been demonstrated in neurological diseases and disorders like the Alzheimer's dementia (Franke and Gaser, 2012), multiple sclerosis (Høgestøl et al., 2019; Cole et al., 2020) schizophrenia (Schnack et al., 2016; Koutsouleris et al., 2014), and other diseases, such as HIV (Petersen et al., 2021; Cole et al., 2017), type 2 diabetes (Franke et al., 2013), and in young adults after premature birth (Hedderich et al., 2021). Evaluating the age gap thus represents an evolving diagnostic biomarker, opening an avenue for researchers to disentangle patterns of brain ageing and associated diseases.

Due to increasing number of publicly available T1w MRI datasets, as well as large private datasets the number of deep learning regression methods for brain age prediction and other brain age studies has soared (Baecker et al., 2021). As opposed to standard machine learning methods, deep learning allows us to train models directly on MR images with minimal preprocessing. Increasingly more accurate age predictions were achieved using various combinations of convolutional neural network (CNN) model architectures, image preprocessing, training strategies, etc. However, due to differences in MRI preprocessing pipelines and software implementations used, it is difficult to disentangle the contribution of methodological innovations from the impact of preprocessing. Furthermore, there is a lack of rigorous statistical analysis to consider the many confounding factors such as the level of grayscale corrections applied,

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

*Corresponding author
lara.dular@fe.uni-lj.si (L. Dular); ziga.spiclin@fe.uni-lj.si (Ž. Špiclin)
ORCID(s): 0000-0002-9947-1501 (L. Dular); 0000-0001-8300-0417 (Ž. Špiclin)

the number of degrees of freedom in T1w to atlas co-registration, software implementation, model architecture and subject/dataset variability, to name a few, rendering an objective comparison between different brain age studies rather difficult.

This paper is organized as follows: a review of related work and our contributions are given in Section 2. In Section 3 we describe the datasets, preprocesing procedures, and DL models for brain age prediction. The evaluation protocol is described in Section 4.1, the experiments with results in Sections 4.2, 4.3 and 4.4. Finally, discussion and conclusion are given in respective Sections 5 and 6.

2. Related Work and Our Contributions

We focus our background review on brain age prediction literature involving the use of deep learning (DL) models. Generally, these are Convolutional Neural Networks (CNN) that input T1w MRIs and are trained to output a scalar value or interval corresponding to the subject's age. Previous studies differ substantially in the number of subject scans, their age span and the nature and level of applied image preprocessing.

Preprocessing pipelines used in brain age studies generally include gray scale enhancement, such as bias field corrections (Lam et al., 2020; Peng et al., 2021; Dufumier et al., 2021; Feng et al., 2020), and registration to a brain atlas. The number of degrees of freedom in registration varied substantially, for instance, and was achieved using rigid (Dartora et al., 2022; Cole et al., 2017), linear (Lam et al., 2020; Peng et al., 2021; Dufumier et al., 2021; Ueda et al., 2019; Huang et al., 2017; Feng et al., 2020) or even nonlinear transforms (Bintsi et al., 2020; Peng et al., 2021; Cheng et al., 2021). Skull stripping that involves extracting the brain from surrounding tissues was also applied in certain studies (Bintsi et al., 2020; Lam et al., 2020; Fisch et al., 2021; Dufumier et al., 2021; Feng et al., 2020; Cheng et al., 2021). In the presence of such preprocessing variations it is difficult to compare study results and disentangle the factors contributing to the accuracy of brain age predictions. A comprehensive study on natural images found that the effect of image preprocessing and augmentation on prediction model performance was greater than the effects of variability in prediction model architecture (Lathuilière et al., 2020), which highlights the need for further research in this area in order to standardize the T1w preprocessing methods for best accuracy of brain age prediction models.

Besides training on the T1w MRIs, brain age models are often trained on segmentation of Gray Matter (GM) and White Matter (WM) structures. Cole et al. (2017) compared models trained on GM segmentations with the model trained on T1w MRI without grayscale corrections, which was rigidly registered to MNI brain atlas. They found that models trained on GM, with mean absolute error (MAE) of 4.16 years, performed better than models trained on T1w with MAE 4.65 and WM images with MAE 5.14. Better performance on GM than on WM segmentations was confirmed by Peng et al. (2021). They further compared models trained on T1w MRIs with bias field correction with both linear and non-linear spatial registration to the MNI brain atlas. The non-linear registration achieved lower MAE of 2.73 years, which was comparable to accuracy achieved by models trained with GM segmentations for linearly registered T1w image with MAE equal to 2.80 years. Finally, Dufumier et al. (2021) reported comparable results of brain age models based on T1w and GM segmentations when testing on same site images, however, the results reported on an independent new-site test set, not used during model training, show preference to models based on GM segmentations.

Differences in T1w preprocessing may arise from the use of different software implementations (Fisch et al., 2021). Related neuroimaging studies show that the measurements of cortical surface thickness differ significantly between pipelines (Kharabian Masouleh et al., 2020; Bhagwat et al., 2021) and reveal a significant discrepancy between the cortical thickness reproducibility metrics (de Fátima Machado Dias et al., 2022). Reasons could also involve T1w MRI resolution variations and contrast-to-noise differences. It seems that the use of GM segmentation for brain age predictions is rather ill-posed and, therefore, this study will focus on preprocessed T1w images as model input. It is yet to be determined if there is a significant effect of the software implementations on brain age prediction even for fairly simple T1w preprocessing operations.

To cut the computational overhead of T1w preprocessing and mitigate potential bias introduced by different software implementations, a recent review paper calls for a further development of brain age from routine MRIs, with minimal preprocessing (Tanveer et al., 2022). The potential and general applicability of such models was already argued by Cole et al. (2017), who proposed one of the first deep learning models for brain age regression. Their model was trained on approximately 2,000 T1w MRIs, with the T1w preprocessing involving only rigid registration to MNI brain atlas, and achieved a MAE of 4.65 years. On a much larger dataset of over 16,000 MRIs using the same minimal T1w preprocessing Dartora et al. (2022) achieved a MAE score of 3.04 years. Further along this line, Fisch et al. (2021) considered minimal T1w preprocessing as applying only skull striping and no spatial registration. Their Residual

Network (ResNet) based model, trained on approximately 10,000 MRIs, achieved MAE of 2.84 years. These recent models seem to achieve competitive results in comparison to previously mentioned models, which were trained on datasets of approximately the same size, but with more extensive T1w preprocessing.

Validation of brain age prediction models for clinical application should involve their performance assessment on new (unseen) site T1 MRI scans. This is a common use case scenario, occurring when applying a pretrained model on new data, possibly preprocessed with a different pipeline. In such scenario, Feng et al. (2020) reported a rather small increase in MAE of 0.15 years, using the same T1w preprocessing on training and test dataset. Multiple other DL studies indicate that this increase (or accuracy deterioration) to be much larger. Jonsson et al. (2019) reported an increase in MAE of about 3 and 5 years on two separate new site datasets. Moreover, Dufumier et al. (2021) showed an increase of MAE by at least 2 years for a wide range of CNN architectures, even for CNNs trained on a large dataset with over 10,000 T1w images.

Drop in brain age prediction accuracy was reported also for models trained on datasets with minimal T1w preprocessing. Dartora et al. (2022) reported a 1 and 3 year increase in MAE on two independent datasets. Fisch et al. (2021) report a 5 year increase on three datasets, prior to applying transfer learning. Therein, the CNN model performed worse than three traditional machine learning models based on explicit feature extraction from T1w MRIs. This increase in MAE therefore seems intrinsically connected to the previously unseen dataset and/or new (unseen) preprocessing procedure and not the model's ability to generalize.

The contributions of this study are the following:

(i) A thorough and reproducible quantitative assessment of the impact of four T1w preprocessing variants on brain age prediction accuracy using four recent model architectures.

(ii) 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, and strategies for its improvement, on a new site dataset and/or new T1w preprocessing approach and software implementation.

3. Datasets and Age Prediction Methodology

3.1. Datasets

For studying the effect of image preprocessing on brain age prediction, we created two datasets: one containing multi-site T1w MRIs for training, validation and testing and the other solely for testing, which contained new unseen site data. All included subjects were healthy individuals from 18 to 95 years old.

The multi-site dataset was gathered from seven publicly available datasets and included a total of 4428 T1w MRIs of healthy subjects. The gathered images were preprocessed using four different preprocessing pipelines and underwent a visual quality check. Images that did not pass the visual quality check 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 overall statistics per dataset are given in Supplementary Table 4. For reproducibility reasons the exact subject IDs included in each split are given in Supplementary materials.

The unseen site dataset was chosen as a subset from the UK Biobank (UKB) dataset. For the purpose of testing, 2815 subjects with a single T1w MRI scan were chosen. The dataset was preprocessed using the same four preprocessing pipelines as multi-site dataset. In addition, a fifth preprocessing pipeline, given by the dataset providers, was used to observe the model's ability to predict not only on previously unseen data, but also on previously unseen preprocessing.

The age distribution of the included T1w subject scans per dataset, and the train/validation/test subsets, is provided in Supplementary Materials (Table 4, Figure 8).

3.2. Image preprocessing pipelines

Four preprocessing pipelines were implemented using a combination of publicly available and in-house software. They differ on the registration methods, the extent of gray scale corrections and the algorithms and software used.

Common to all pipelines, the input T1w image was first converted to the Nifti format. Common reference space was the MNI152 nonlinear atlas, version 2009c (Fonov et al., 2009), with size $193 \times 229 \times 193$ and spacing 1 mm^3, to which the input T1w image was sinc resampled. All subsequently described image registration steps and the sinc resampling were performed using the publicly available NiftyReg software[1] (Modat et al., 2014).

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 (Manjón et al., 2010).

Aligned with Cole et al. (2017), the first pipeline, denoted **RIG**, performed rigid registration of the denoised T1w image into the MNI152 atlas space. To improve registration accuracy, intensity inhomogeneity correction (without mask) was applied to the denoised image using N4 algorithm [3] (Tustison et al., 2010), 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: *1)* intensity windowing, in which intensity values above 99-th and below 5-th percentile were set to corresponding limiting values, and *2)* intensity inhomogeneity correction using the N4 algorithm on 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.

The fourth pipeline, **Fs+FSL**, again included gray scale corrections and affine registration, but was based on commonly used software tools FreeSurfer[4] and FSL (FMRIB Software Library) [5] (Jenkinson et al., 2012). The raw T1w images were preprocessed using the first four preprocessing stages of the FreeSurfer's cortical reconstruction `recon-all` pipeline, which entailed non-parametric non-uniform intensity normalization (N3), MNI305 atlas using the MINC program mritotal (Collins et al., 1994) and intensity normalization that attempts to correct for fluctuations in intensity and assures all voxels are scaled so that the mean intensity of the white matter is 110 (Laboratory for Computational Neuroimaging and Athinoula A. Martinos Center for Biomedical Imaging., 2022). In order to ensure consistency among all preprocessing pipelines, we also applied a registration step to the T1w images processed using Freesurfer. Specifically, we used FSL for registration to the MNI152 nonlinear atlas, version 2009c Fonov et al. (2009). This step was necessary as Freesurfer preprocessing includes a registration to a different brain atlas than the other pipelines. During this step, the T1w images were reoriented to RAS orientation and registered into the MNI152 atlas using FLIRT (Jenkinson et al., 2002). The registration process was executed in two steps: first, a rigid registration with 6 degrees of freedom was applied, followed by a linear registration with 12 degrees of freedom.

3.2.1. Adapting UKB preprocessed data

An additional, fifth variant of preprocessed T1 MRIs from the UKB dataset, described in detail by Smith et al. (2020), was obtained from the UKB dataset providers. Namely, from the UKB dataset we included raw T1w defaced images in subject image space, as well as the preprocessed T1w images and corresponding linear transformation matrices that register the raw T1w image to MNI152 nonlinear 6th generation atlas space (Grabner et al., 2006). Since the above four preprocesing procedures involved registration to 7th generation MNI152 atlas, an additional common linear registration between 6th and 7th generation atlas spaces was applied, to assure all images were in the same space.

Each original defaced T1w image was first resampled to the MNI152 nonlinear 6th generation atlas (Grabner et al., 2006) using FSL FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002) and the provided linear transformation matrix, and then linearly registered to the MNI152 nonlinear 7th generation MNI atlas (version 2009, our target space) (Fonov et al., 2009) and resampled using 3rd order interpolation. The linear transformation matrix between the two MNI spaces was pre-computed using FSL's FLIRT.

3.2.2. Adapting image size

In all resulting preprocessed T1w MRI images we removed the non-informative empty space around the head by cropping to size $157 \times 189 \times 170$ about the image center.

[1]NiftyReg Software `http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg`
[2]Adaptive non-local means denoising implementation: `https://github.com/djkwon/naonlm3d`
[3]N4 bias field correction: `https://manpages.debian.org/testing/ants/N4BiasFieldCorrection.1.en.html`
[4]Freesurfer: `https://surfer.nmr.mgh.harvard.edu/`
[5]FSL (FMRIB Software Library): `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki`

Further, the image size as input to the models was adapted during the augmentation (3.3.2). Namely, 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 by a factor of 2 using sinc resampling and cropped to size $95 \times 79 \times 78$.

3.3. 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.

3.3.1. Model architecture

Model 1, proposed by Cole et al. (2017), was a convolutional CNN trained on full resolution 3D T1w MRIs. **Model 2**, proposed by Huang et al. (2017), was trained on 2D images by taking 15 equidistantly sampled axial slices as input channels. **Model 3**, proposed by Ueda et al. (2019), was trained on downsampled T1w MRIs. Finally, **Model 4**, proposed by Peng et al. (2021), was a fully convolutional model trained on full resolution 3D images that reported one of best results for brain age prediction among the CNN models. The architectures of the four models are depicted in Figure 1.

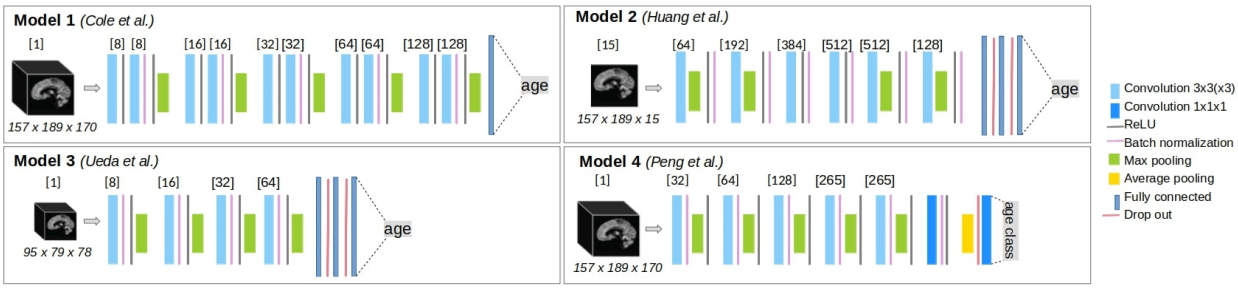

Figure 1: Architecture of the four reimplemented CNN models for brain age prediction.

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 his or her T1 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.

3.3.2. Model Training

Hyperparameter tuning. The learning rate (LR) and batch size hyperparameter values for each model were chosen based on a wide grid search, which was set around the proposed values in corresponding original papers. For instance, tested LR values were 10^{-2}, 10^{-3}, 10^{-4}, $5 \cdot 10^{-5}$, 10^{-5}, and 10^{-6}. The batch size for Models 2 and 3 was set to 4, 8, 16, 32 and 64. Due to GPU constraints we trained Model 1 with batch size 4, 8, 16 and 24 and Model 4 with batch size 4 and 9. All tested hyperparameter combinations and their results are given in Supplementary Figure 9.

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 both mean squared error (MSE) loss and L1 loss for multiple hyperparameter values. Due to overall better performance and stability of training, all three models were trained with L1 loss. Model 4, defined as a classification model, was trained with Kullback-Leibler divergence (KLD) loss function.

Optimizer. We used the SGD algorithm with momentum 0.9 as proposed in three out of four studies (Cole et al., 2017; Peng et al., 2021), keeping the LR 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.

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 necessary for classification Model 4, but not for the other three models[6].

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. (2020) 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.

3.4. Postprocessing

The following postprocessing steps, common in brain age prediction literature, were applied for improving prediction performance and reliability.

Bias correction. As a result of regression dilution, a systematic over- and under-estimation of brain age on lower and upper end of the dataset age span can be observed. To alleviate this phenomenon many researches apply some form of post-hoc correction of the predictions in form of (linear) bias correction on training or validation set (Lange et al., 2019; Peng et al., 2021; Cole et al., 2017; Smith et al., 2019; Cheng et al., 2021; Dunås et al., 2021).

We applied a linear correction by fitting a regression line $y = \beta_1 \hat{y} + \beta_0$ on the validation set, where y denotes true and $\hat{y}$ predicted value. The estimated coefficients β_0 and β_1 on the validation set, were then used for correcting the predicted brain age on the test set as $y' = \beta_1 \hat{y} + \beta_0$.

Model ensembling. Model ensembling was shown effective in reducing the MAE values, both when combining model outputs obtained from single (Peng et al., 2021; Dufumier et al., 2021; Levakov et al., 2020; Cheng et al., 2021) or multiple preprocesing pipelines (Peng et al., 2021; Kuo et al., 2021).

To avoid reporting the results of a single (possibly lucky) 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. In all experiments, model ensembling was applied after bias correction.

4. Experiments and results

The impact of T1-w MRI image preprocessing on the accuracy of brain age predictions using the four CNN models were studies in three scenarios shown in Figure 2: *1)* tested on the same dataset and preprocessing used during model training (**Section 4.2**), *2)* tested on a new unseen dataset but preprocessed in the same way as training dataset, *3)* tested on a new unseen dataset, preprocessed differently than training dataset.

4.1. Evaluation Protocol

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

[6]The significance of weighted training was statistically confirmed: reduction in AE was statistically significant for a subgroup of subjects over the age of 80 years.

	Section 4.2	**Section 4.3**	**Section 4.3**
Training dataset	Multi-site dataset	Multi-site dataset	Multi-site dataset
Test dataset	Multi-site dataset	New dataset	New dataset
Test dataset preprocessing	Same as training	Same as training	New preprocessing

Figure 2: Overview of the tested brain age train and test scenarios.

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 true age and y_i' predicted age of i-th subject. We also report mean error (ME):

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

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.

4.1.1. Statistical Analysis

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, bias correction). Additional fixed effects were included depending on the research question at hand.

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 all LMEM we reported regression coefficients and their 95% confidence intervals. Results of LMEM analysis were supported by the ANOVA test declaring statistical significance for $p < 0.01$. 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 method.

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.

4.2. Effect of image preprocessing

Our goal is to evaluate the impact of the particular choice of image preprocessing for various CNN architectures, described in respective Sections 3.2 and 3.3. 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 initializations. Brain age predictions were obtained as the average age predictions of five models trained with different random weight initialization.

	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 ± 4.09	3.18 ± 2.62	-0.45 ± 4.14	$\underline{3.13 \pm 2.75}$	-0.34 ± 3.86	$\underline{\mathbf{2.97 \pm 2.48}}$	-0.17 ± 3.99	$\underline{3.08 \pm 2.54}$
Model 2	-0.22 ± 5.68	4.32 ± 3.69	-0.01 ± 5.43	4.14 ± 3.50	0.20 ± 4.74	$\underline{\mathbf{3.53 \pm 3.16}}$	0.07 ± 5.55	4.26 ± 3.55
Model 3	-0.06 ± 4.41	3.45 ± 2.74	0.01 ± 4.39	3.39 ± 2.79	-0.19 ± 3.9	$\mathbf{3.14 \pm 2.32}$	0.11 ± 4.32	3.42 ± 2.64
Model 4	0.05 ± 4.26	$\underline{3.17 \pm 2.85}$	0.06 ± 4.27	$\mathbf{3.15 \pm 2.88}$	0.18 ± 4.15	3.17 ± 2.68	0.22 ± 4.54	3.42 ± 2.98

Table 1: Multi-site test set results for all 16 combinations of preprocessing pipelines and 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.

The model accuracy metrics are presented in Table 1. Models 1, 2 and 3 trained on data with the RIG preprocessing resulted in higher MAE values than with other preprocessing pipelines. With Model 4 the metric values were comparable among the RIG, RIG+GS and AFF+GS pipelines with MAE of approximately 3.15 years.

Compared to the RIG pipeline, the RIG+GS included gray scale correction steps, however, resulted in only a marginal decrease in MAE for 3D Models 1, 3 and 4. Slightly larger decrease in MAE was observed for 2D Model 2. Nonetheless, this difference was not significant for any of the models.

The largest decrease in MAE can be observed when switching between rigid and affine registration; for instance, as much as 0.83 years for Model 2. With Model 1 the reduction of MAE was 0.16 years, to 2.97 years, which was the best MAE score reported in this study.

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 are statistically significant. We increased the LMEM complexity by including the interaction of the fixed factors, however the interaction terms were not statistically significant. Since the main effects are statistically significant and the interaction is theoretically meaningful, the interaction was included in the final model despite not being statistically significant. The regression 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 Figure 3. The analysis revealed that the pairwise differences were statistically significant only between 2D Model 2 and the three 3D Models 1, 2, and 3. The AE of Model 2 was found to be significantly different from those of Models 1, 3, and 4, except for the case of datasets that underwent AFF+GS preprocessing. Although all models demonstrated improved performance on datasets that underwent affine registration, these differences were significant only for Model 2. Specifically, the differences were significant between AFF+GS and RIG, as well as AFF+GS and Fs+FSL.

4.3. Performance on unseen data

In this experiment, we evaluated the performance of 80 trained models on unseen data. In general, new data may come from a different MRI scanner or have undergone different preprocessing than the data used to train the models. However, for the purpose of this experiment, we assumed that the unseen data had been preprocessed in the same way as the training data, which is a common scenario in practice. For example, this may occur when applying a pre-trained model to new data or when predicting brain age on a dataset from a scanner or location that was not available during model training.

To evaluate the performance of the models on the unseen data, we applied them to the 2815 T1w scans from the UKB dataset without any additional training. To correct for bias, we used precomputed β_0 and β_1 values from the multi-site validation dataset and averaged the results across five models with different weight initializations. This resulted in a total of 16 predictions for each T1w image, which served as our baseline.

Assuming the predictions offset on new dataset is systematic and can be reduced with a simple bias correction (Franke and Gaser, 2012; Dular and Špiclin, 2021), we comparatively evaluated *1)* the baseline predictions as described above, and *2)* further bias correction computed on a subset of UKB dataset. Here, UKB specific β_0 and β_1 for each trained model were computed on approximately 10% randomly chosen subjects and applied on the remaining 90% of the dataset. The predictions were then averaged across five models with different weight initialization.

For estimating the influence of preprocessing pipeline on model performance on unseen data we fit a LMEM model with architecture, preprocessing, BC approach, their two-way and three-way interactions as fixed effects, and subject ID

		Model 1				Model 2				Model 3				Model 4			
		RIG	RIG+GS	AFF+GS	Fs+FSL	RIG	RIG+GS	AFF+GS	Fs+FSL	RIG	RIG+GS	AFF+GS	Fs+FSL	RIG	RIG+GS	AFF+GS	Fs+FSL
Model 1	RIG		0.05	0.21	0.10	-1.14	-0.96	-0.35	-1.08	-0.27	-0.21	0.04	-0.24	0.01	0.03	0.01	-0.24
	RIG+GS	-0.05		0.16	0.05	-1.19	-1.01	-0.40	-1.13	-0.32	-0.26	-0.01	-0.29	-0.03	-0.02	-0.04	-0.29
	AFF+GS	-0.21	-0.16		-0.11	-1.35	-1.17	-0.56	-1.28	-0.48	-0.42	-0.17	-0.45	-0.19	-0.18	-0.20	-0.45
	Fs+FSL	-0.10	-0.05	0.11		-1.24	-1.06	-0.45	-1.18	-0.37	-0.31	-0.05	-0.33	-0.08	-0.07	-0.09	-0.34
Model 2	RIG	1.15	1.19	1.35	1.25		0.18	0.79	0.06	0.87	0.93	1.18	0.90	1.16	1.17	1.15	0.90
	RIG+GS	0.96	1.01	1.17	1.06	-0.18		0.61	-0.11	0.69	0.75	1.00	0.72	0.98	0.99	0.97	0.72
	AFF+GS	0.35	0.40	0.56	0.45	-0.79	-0.61		-0.72	0.09	0.14	0.34	0.11	0.37	0.38	0.36	0.11
	Fs+FSL	1.08	1.13	1.28	1.18	-0.06	0.11	0.72		0.81	0.87	1.12	0.84	1.09	1.10	1.09	0.84
Model 3	RIG	0.27	0.32	0.48	0.37	-0.87	-0.69	-0.09	-0.81		0.06	0.31	0.03	0.28	0.30	0.28	0.03
	RIG+GS	0.21	0.26	0.42	0.31	-0.93	-0.75	-0.14	-0.87	-0.06		0.25	-0.03	0.22	0.24	0.22	-0.03
	AFF+GS	-0.04	0.01	0.17	0.05	-1.18	-1.00	-0.34	-1.12	-0.31	-0.25		-0.28	-0.02	-0.15	-0.03	-0.28
	Fs+FSL	0.24	0.29	0.45	0.33	-0.90	-0.72	-0.11	-0.84	-0.03	0.03	0.28		0.25	0.27	0.25	0.00
Model 4	RIG	-0.01	0.03	0.19	0.08	-1.16	-0.98	-0.37	-1.09	-0.28	-0.22	0.02	-0.25		0.01	-0.00	-0.25
	RIG+GS	-0.03	0.02	0.18	0.07	-1.17	-0.99	-0.38	-1.10	-0.30	-0.24	0.15	-0.27	-0.01		-0.02	-0.27
	AFF+GS	-0.01	0.04	0.20	0.09	-1.15	-0.97	-0.36	-1.09	-0.28	-0.22	0.03	-0.25	0.00	0.02		-0.25
	Fs+FSL	0.24	0.29	0.45	0.34	-0.90	-0.72	-0.11	-0.84	-0.03	0.03	0.28	-0.00	0.25	0.27	0.25	

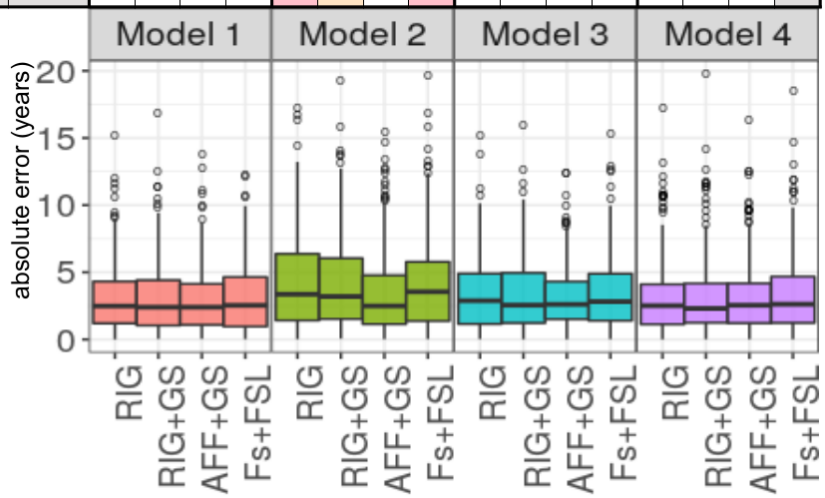

Figure 3: Results of LMEM post-hoc pairwise statistical tests for all image preprocessing and model architecture combinations. The color of each square marks statistical significance: red for $p < 0.001$, orange for $0.01 \leq p < 0.001$, yellow for $0.05 \leq p < 0.01$ and white for $p > 0.05$ (not significant).

as random effect. ANOVA test confirmed that this model explained more variability than model with no interactions and model with two-way interactions ($p < 0.001$). ANOVA test of effects show all main effects, their two-way, and three-way interactions as statistically significant ($p < 0.001$), except for interaction between bias correction and model architecture ($p = 0.96$), meaning the results do not provide the evidence that a difference in MAE between architectures is affected by applying bias correction on UKB dataset. Detailed results are presented in Supplementary Table 8.

Table 2 and Figure 4 show the mean ensemble MAE values for the 16 combinations of preprocessing and model architecture. The baseline MAE values for the UKB dataset range from 5.12 years for the 2D Model 2 with RIG

preprocessing to 3.66 years for Model 1 with Fs+FSL preprocessing. Additionally, the 3D models trained on datasets with affine registration had a smaller bias as measured by the ME. Notably, the 2D model before BC often had the smallest bias, but also had the largest standard deviation for most preprocessing pipelines.

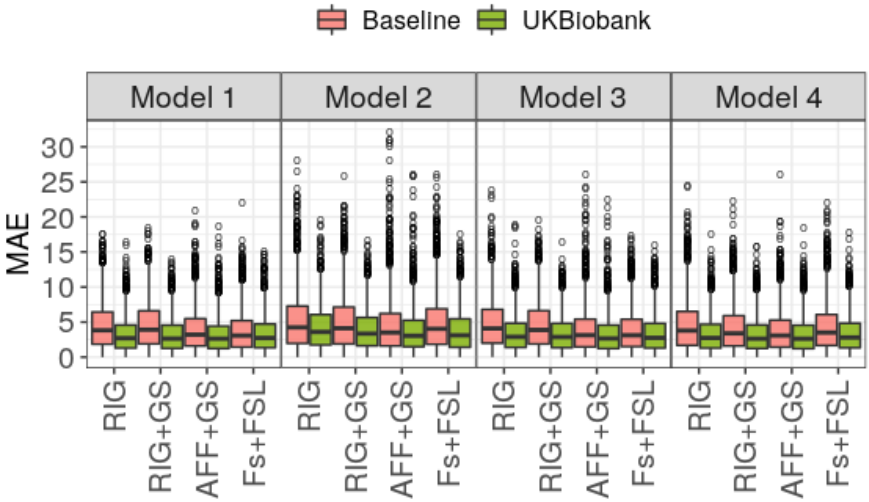

Figure 4: Absolute error (AE) in years on UKB dataset preprocessed the same way as training data. Comparison of baseline model without bias correction (pink) and with bias correction on a subset of UKB subjects (green).

Correcting for bias on UKB dataset results in a 0.94 year decrease in MAE across all datasets and preprocessing procedures. BC marginal means conditional to preprocessing and model interaction are significant for all architecture and preprocessing combinations ($p < 0.001$). BC reduced most of the bias, however even after, the models slightly underestimate age from approximately -0.02 to -0.54 years (Table 2). BC reduced not only MAE and ME, but also their standard deviation across all combinations, to values comparable to those in Table 1. We should however note, that the age in the UKB dataset spans half the age interval of multi-site dataset.

Out of preprocessing procedures, more intensive preprocesing procedures, Fs+FSL and AFF+GS resulted in the lowest MAE before BC, however after BC, AFF+GS yielded the best MAE results. Before applying BC, for all four preprocessing procedures full resolution 3D Models 1 and 4 outperformed Models 2 and 3, trained on reduced information. This pattern remained the same after BC, however Model 1 performed best after BC for all four preprocessing procedures with an overall best result of Model 1 with AFF+GS 3.11 ± 2.4 after BC.

	Bias 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	**BL**	-3.43 ± 4.25	$\underline{4.43 \pm 3.19}$	-3.57 ± 4.29	4.53 ± 3.26	-2.25 ± 4.28	3.84 ± 2.94	-1.12 ± 4.49	$\underline{\mathbf{3.66 \pm 2.84}}$
	UKB	-0.35 ± 3.98	$\underline{3.18 \pm 2.42}$	-0.45 ± 3.91	$\underline{3.14 \pm 2.37}$	-0.44 ± 3.9	$\underline{\mathbf{3.11 \pm 2.40}}$	-0.28 ± 4.11	$\underline{3.28 \pm 2.50}$
Model 2	**BL**	-0.46 ± 6.48	5.12 ± 4.01	0.13 ± 6.27	4.95 ± 3.85	0.01 ± 5.83	$\mathbf{4.41 \pm 3.81}$	-1.39 ± 6.21	4.94 ± 4.01
	UKB	-0.29 ± 5.19	4.19 ± 3.07	-0.23 ± 4.84	3.91 ± 2.86	-0.38 ± 4.85	$\mathbf{3.75 \pm 3.10}$	-0.07 ± 4.82	3.79 ± 2.98
Model 3	**BL**	-3.53 ± 4.67	4.75 ± 3.43	-3.14 ± 4.73	4.57 ± 3.36	-1.66 ± 4.67	3.82 ± 3.16	-0.82 ± 4.73	$\mathbf{3.80 \pm 2.93}$
	UKB	-0.54 ± 4.22	3.39 ± 2.58	-0.49 ± 4.15	3.35 ± 2.51	-0.28 ± 4.11	$\mathbf{3.20 \pm 2.59}$	-0.28 ± 4.21	3.34 ± 2.58
Model 4	**BL**	-2.81 ± 5.07	4.54 ± 3.60	-1.68 ± 5.00	$\underline{4.14 \pm 3.27}$	-0.13 ± 4.74	$\underline{\mathbf{3.72 \pm 2.95}}$	0.89 ± 5.34	4.24 ± 3.36
	UKB	-0.29 ± 4.10	3.26 ± 2.51	-0.22 ± 4.00	3.17 ± 2.45	-0.30 ± 3.97	$\mathbf{3.13 \pm 2.47}$	-0.02 ± 4.26	3.38 ± 2.60

Table 2: MAE on unseen dataset (UKB), preprocessed in the same manner as multi-site training data. Results are presented for 16 models and preprocessing combinations, with and without additional BC on UKB dataset. 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.
Abbreviations: *BL* – baseline results, without bias correction; *UKB* – results with bias correction on UKB dataset.

Since BC on UKB dataset significantly reduced MAE, we further limited ourselves to these results to study the effect of preprocessing procedure for different model architectures. We fit a LMEM model with preprocessing, model

architecture and their interaction as fixed effects and subject ID as random effect (Supplementary Table 9). ANOVA test showed that both main effects and their interaction was statistically significant ($p < 0.001$).

	Model 1				Model 2				Model 3				Model 4			
	R	R+G	A+G	F+F	R	R+G	A+G	F+F	R	R+G	A+G	F+F	R	R+G	A+G	F+F
R		0.04	0.08	-0.09		0.29	0.44	0.40		0.04	0.19	0.05		0.09	0.13	-0.12
R+G	-0.04		0.03	-0.14	-0.29		0.15	0.11	-0.04		0.14	0.01	-0.09		0.04	-0.21
A+G	-0.08	-0.03		-0.17	-0.44	-0.15		-0.04	-0.19	-0.14		-0.13	-0.13	-0.04		-0.25
F+F	0.09	0.14	0.17		-0.40	-0.11	0.04		-0.05	-0.01	0.13		0.12	0.21	0.25	

Figure 5: The pairwise difference in marginal means between preprocessing procedures conditional on model architecture.The preprocessing procedure of test data was the same as preprocessing procedure of train data. 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).

Figure 5 shows the pairwise difference in marginal means of preprocessing conditional on model architecture for the above LMEM and their statistical significance. We can indeed observe that the marginal MAE of AFF+GS preprocessing is lower than marginal means of all other preprocessing values, however this difference is not always significant. Interestingly, for Models 1 and 4 trained on full resolution images, this difference is most significant between AFF+GS and Fs+FSL preprocessing that both include similar grayscale corrections and affine registration. Moreover, models trained on Fs+FSL preformed worse that models trained on preprocessing with only rigid registration, however this difference was not significant. Due to the use of partial (subsampled) imaging information in Models 2 and 3 the differences between AFF+GS preprocesing were larger and statistically significant.

Finally, Model 2, trained on 2D image datasets with affine registration, outperformed those with rigid registration, which shows to its sensitivity to spatial information.

4.4. Performance on unseen data with new image preprocessing

We further considered the cumulative effect of dataset not used during model training, additionally with different image preprocessing as dataset used during model training, as given by the dataset provider, described in Section 3.2.1. Without additional training we predicted the age for all 80 trained models. The model predictions were bias corrected using β_0 and β_1 computed on the validation set of the multi-site T1w dataset across five models with different weight initialization, which resulted in 16 predictions per each T1w image (baseline).

Similarly as in previous experiment we comparatively evaluated *1)* the baseline predictions, and *2)* predictions using bias correction on UKB dataset, by fitting a regression line on the same 10% of subjects and used the estimated coefficients for correcting the predicted brain age on the rest of the data.

The MAE and ME metrics of predictions of the 16 mean ensembles were compared between the two bias correction approaches. We fit a LMEM with absolute error as dependent variable, subject ID as random effect, and bias correction approach, image preprocessing, model architecture and their interactions as fixed effects, computed the ANOVA test and the 95% CI intervals.

For estimating the influence of preprocessing pipeline on performance on unseen dataset with new preprocessing we fit a LMEM model with BC, architecture, preprocessing, their two-way and three-way interactions as fixed effects, and subject ID as random effect. ANOVA test confirmed that the model explained more variability then and models with no interactions and model with two way interactions ($p < 0.001$). ANOVA test of effects show all main effects, their two way and three way interactions as statistically significant ($p < 0.001$). Details are provided in Supplementary Table 10.

The mean ensemble MAE values for all 16 preprocessing and model architecture combinations are presented in Table 3, while Figure 6 shows box-plots of the MAE values for baseline with and without BC on UKB dataset.

In baseline results the predicted age is generally underestimated on the unseen UKB T1w test scans, with ME as low as -17.09 years for Model 2 on RIG+GS dataset and only 0.46 years for Model 4 with Fs+FSL preprocessing. This may be expected considering the smaller age span of the UKB population versus the multi-site train set population (cf. Supplementary Table 4). The large MAE values close to 40 years were observed for Model 2. The model exhibits

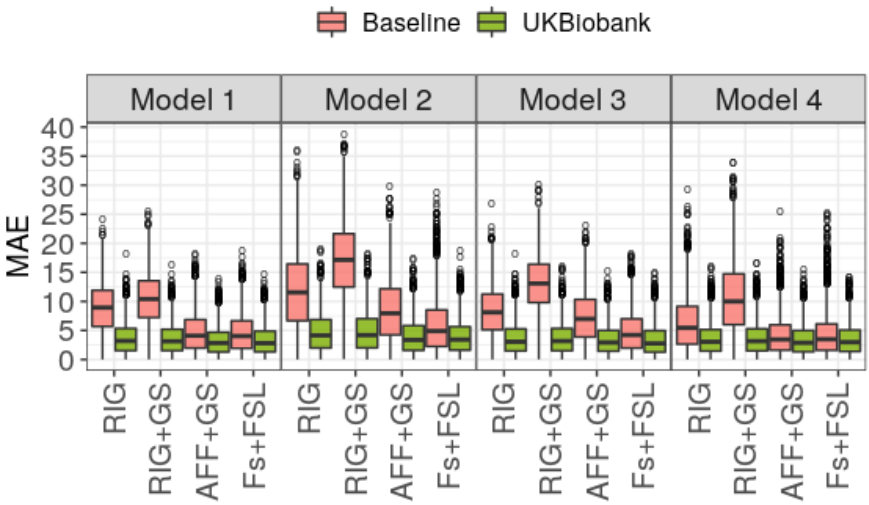

Figure 6: Mean absolute error (MAE) in years for all preprocessing and model architecture combinations for baseline model (without bias correction) and with bias correction on a subset of UKB subjects.

large bias and errros across all four image preprocessing procedures with MAE ranging from 17.11 years for RIG+GS preprocessing to 5.97 years for Fs+FSL preprocessing.

All 3D models perform substantially better than 2D model across all preprocessing procedures. There is a large difference in performance between models trained on datasets with affine registration and rigid registration, for which the MAE approximately doubled. For instance Model 1 trained on RIG+GS dataset performed with a MAE of 10.39 years and 4.57 years on Fs+FSL preprocessing. Best baseline results for three out of four models were achieved on Fs+FSL pretrained model, which can be attributed to the fact that the same software were used for preprocessing of UKB dataset (Section 3.2.1).

	Bias	RIG		RIG+GS		AFF+GS		Fs+FSL	
	Corr.	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)	ME (sd)	MAE (sd)
Model 1	**BL**	-8.76 ± 4.64	8.89 ± 4.38	-10.33 ± 4.76	10.39 ± 4.62	-3.36 ± 4.68	4.70 ± 3.34	-3.14 ± 4.67	$\mathbf{4.57 \pm 3.27}$
	UKB	-0.15 ± 4.59	3.70 ± 2.71	-0.28 ± 4.49	3.64 ± 2.65	-0.44 ± 4.03	$\mathbf{3.26 \pm 2.41}$	-0.41 ± 4.18	3.35 ± 2.53
Model 2	**BL**	-11.64 ± 7.17	11.91 ± 6.71	-17.09 ± 6.77	17.11 ± 6.71	-8.09 ± 6.04	8.55 ± 5.38	-3.03 ± 7.01	$\mathbf{5.97 \pm 4.76}$
	UKB	-0.65 ± 5.71	4.69 ± 3.33	-0.63 ± 5.81	4.76 ± 3.39	-0.31 ± 4.97	$\mathbf{3.98 \pm 2.99}$	-0.18 ± 4.96	$\mathbf{3.98 \pm 2.97}$
Model 3	**BL**	-8.09 ± 4.55	8.24 ± 4.28	-13.06 ± 4.89	13.08 ± 4.83	-6.99 ± 4.81	7.29 ± 4.33	-3.05 ± 5.06	$\mathbf{4.78 \pm 3.47}$
	UKB	-0.33 ± 4.47	3.59 ± 2.68	-0.37 ± 4.60	3.71 ± 2.74	-0.52 ± 4.25	3.44 ± 2.53	-0.36 ± 4.23	$\mathbf{3.36 \pm 2.60}$
Model 4	**BL**	-5.35 ± 5.81	6.37 ± 4.68	-10.57 ± 6.46	10.74 ± 6.17	-1.12 ± 5.41	$\mathbf{4.26 \pm 3.51}$	-0.46 ± 5.76	4.39 ± 3.76
	UKB	-0.38 ± 4.51	3.60 ± 2.74	-0.57 ± 4.58	3.69 ± 2.77	-0.49 ± 4.27	$\mathbf{3.43 \pm 2.59}$	-0.19 ± 4.39	3.52 ± 2.64

Table 3: Mean ensemble MAE values for 16 preprocessing and model architecture combinations, with and without BC on UKB dataset with new preprocessing procedure. 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.

Bias correction on the UKB subset computed on 10% of subjects notably improved the mean predictions, driving the ME values for all combinations of image preprocessing and model architectures between -0.65 and -0.18 years, however its standard deviation was up to 0.5 years higher compared to results in Section 4.3. Pairwise comparison of BC marginal means conditional to preprocessing and model interaction is significant for all architecture and preprocessing combinations ($p < 0.001$).

Limiting ourselves to the results with BC on UKB, we fit a LMEM model with preprocessing, model architecture and their interaction as fixed effects and subject ID as random effect (Supplementary Table 11) in order to study the interactions between new preprocessing and processing of pretrained model. ANOVA test show both main effect and their interaction as statistically significant ($p < 0.001$).

	Model 1				Model 2				Model 3				Model 4			
	R	R+G	A+G	F+F	R	R+G	A+G	F+F	R	R+G	A+G	F+F	R	R+G	A+G	F+F
R		0.07	0.44	0.35		-0.07	0.71	0.71		-0.11	0.15	0.24		-0.08	0.17	0.09
R+G	-0.07		0.38	0.29	0.07		0.78	0.79	0.11		0.26	0.35	0.08		0.25	-0.17
A+G	-0.44	-0.38		-0.09	-0.71	-0.78		0.01	-0.15	-0.26		0.09	-0.17	-0.25		-0.08
F+F	-0.35	-0.29	0.09		-0.71	-0.79	-0.01		-0.24	-0.35	-0.09		-0.09	-0.17	0.08	

Figure 7: The pairwise difference in marginal means between preprocessing procedures conditional on model architecture. The preprocessing procedure of test data differed from preprocessing procedure of train data. 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).

Figure 7 shows that the pairwise difference in marginal means of preprocessing, conditional on model architecture for LMEM is statistically significant for almost all pairs of models with rigid registration and models with affine registration. This improvement ranged from 0.17 years for Model 4 to almost 0.8 years for 2D Model 2. It seems that despite bias correction on UKB, the similarity between preprocessing of new data and preprocessing of training data remained significant for the end prediction.

The difference between the two preprocessing procedures with affine registration, AFF+GS and FS+Fsl, was not significant for any model. Again, there was no significant improvement of model predictions for gray scale corrections of preprocessing procedures RIG and RIG+GS.

5. 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. Each model was initialized and trained five times and we reported the mean values of the five model predictions across all model architecture and T1w preprocessing combinations. We further considered predicting on a new T1w image dataset, not seen during model training, which was preprocessed the same as the training dataset, or in a different manner using different operators or software implementation.

The point estimates of brain age accuracy like the MAE, which are usually reported in brain age literature, need to be statistically evaluated to enable one to draw generalizing conclusions. For this purpose we used the linear mixed-effects models (LMEMs), as they allow us to account for repeated measures on a subject level by including the subject ID as a random effect. As our results show, despite observed difference in point estimates of the MAE values, the difference may not be statistically significant. For instance, when comparing the performances of Models 1 and 3 obtained with the AFF+GS preprocessing (cf. Table 1), the seemingly relevant difference in MAE values of about 0.2 years was not statistically significant (cf. Figure 3).

5.1. Impact of T1-weighted image preprocessing and model architecture

When comparing the effect of T1w preprocessing with respect to the four T1w preprocessing pipelines, a slightly higher brain age prediction accuracy (i.e. low MAE) across all models was observed for the AFF+GS preprocessing pipeline. Interestingly, the inclusion of gray scale correction into the pipeline, such as denoising and intensity inhomogeneity correction did not improve MAE, but was needed for accurate (linear) image registration used in our preprocessing pipelines.

Despite the high similarities between the T1w preprocessing pipelines, we noticed a significant difference between models trained using common software such as FSL and FreeSurfer and AFF+GS pipeline. This shows that even when training brain age prediction models on the same source data, but with different implementations of T1w preprocessing software, the obtained results may not be directly comparable.

The results cannot be directly compared even to the results in original papers, in which the CNN models evaluated in this study were proposed, due to the differences in size and age structure of the training dataset. Interestingly, the

MAE of Model 1 reported herein was 1 year lower than the MAE reported by Cole et al. (2017), despite the fact that the T1w preprocessing (RIG), structure and size of training set were similar. We attribute the improvement partially to the mean ensembling and largely to extensive hyperparameter tuning. In general, all four model architectures performed considerably better if T1w preprocessing involved linear (affine) registration, i.e. the AFF+GS and Fs+FSL. This indicates the importance of good spatial normalization of the input T1w scans, which eliminates the inter-subject variance due to head size differences and MRI-acquisition related geometric artifacts. However, pairwise comparison was only marginally statistically significant between the AFF+GS and RIG, and AFF+GS and Fs+FSL T1w preprocessing pipelines.

Obtaining better results with higher degree of freedom spatial image registration supports the hypothesis of Dartora et al. (2022), who attribute worse results of their model compared to previous publications to the lower degree or complexity of the T1w preprocessing. Further, the results are in inline with the study by Peng et al. (2021), wherein the T1w preprocessing procedures including either linear or non-linear registration were compared, resulting in slight favor of the latter.

In their review Tanveer et al. (2022) discuss computational complexity and call on research community to further focus on the 2D CNN brain age prediction models. Our results show a statistically significant inferiority of the implemented 2D model versus all tested 3D models, regardless of the T1w preprocessing applied. This finding is in line with Feng et al. (2020), 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 reimplementations of the 3D models, but need to introduce a methodological improvement, like for instance Jönemo et al. (2022), predicting age from 2D projections of the 3D MRI volumes.

The final aspect when comparing different preprocessing procedures is the computational complexity. To use the brain age as a prognostic biomarker in clinical practice, the T1w brain MRI should be processed in a reasonably short time. Despite longer executions times of some presented preprocessing pipelines (1.5 to 16.5 minutes), substantial gains seem possible with the use of GPU (re-)implementations. The long training times of the DL models do not seem a limitation, as they are performed off-line, while the model inference time per each brain age prediction is typically in the order of a few seconds and thus negligible compared to the T1w preprocessing time.

5.2. Performance on new unseen scanner data

In contrast to the differences in MAE predictions on multi site-dataset, where only marginally significant differences were observed in part due to the smaller sample size, the differences between models and preprocessing procedures were apparent when inferring on the new unseen data. Despite a larger test sample we again note that the MAE difference obtained on T1w images from two preprocessing procedures with affine registration and different grayscale corrections was minimal and not statistically significant for any of the 3D brain age models.

Among the T1w preprocessing procedures evaluated, those that were more extensive (Fs+FSL and AFF+GS) exhibited the lowest brain age prediction errors before bias correction. However, after applying bias correction, the AFF+GS consistently produced the best results in terms of the smallest MAE and its standard deviation, regardless of the brain age model architecture. This difference was statistically significant for all brain age models except the Model 2.

Despite recent efforts to use minimally preprocessed T1w images (Dartora et al., 2022; Fisch et al., 2021) and calls for further development of models on routine MRIs (Tanveer et al., 2022), our findings suggest that using more extensive T1w preprocessing can improve the prediction accuracy of brain age models even on datasets obtained from a new site. The sensitivity to the type of spatial registration of the T1w image to the brain atlas space was particularly crucial for the 2D model, which was trained on 15 axial slices and performed significantly better on datasets with the affine registration, resulting in an improvement in MAE of 0.4 years.

However, it is worth noting that other factors, such as the size and characteristics of the dataset, may also influence the brain age prediction accuracy. For example, the models tend to underestimate the age of the subjects in the UKB dataset, both before and after bias correction, which may be due to the higher age of the individuals in the training dataset. Additionally, the observed MAE values of 2.97 years for the multi-site test dataset and 3.11 for the new-site dataset were comparable, which may be partially attributed to the smaller age range of the subjects in the UKB dataset. However, the MAE is unlikely to increase proportionally with the increase of age range for adult datasets as assumed by Cole et al. (2019). For instance, in an experiment conducted by Peng et al. (2021), the Model 2 was trained on UKB and again on dataset with age range from 17 to 90 years of somewhat similar sizes (i.e. 2600 and 2200 subjects) and achieved MAE values of 2.76 years and 2.9 years, respectively.

5.3. Transferability of model on dataset with new preprocessing

Differences in the level of applied T1w preprocessing between the training and test set played a crucial role when predicting brain age on new unseen scanner data. For instance, the values of MAE obtained with RIG and/or RIG+GS pipelines were more than double the values of the MAE obtained with the Fs+FSL and/or AFF+GS pipelines. The MAE with Fs+FSL was only about 1 year worse than the MAE obtained when the same T1w preprocessing was applied prior to bias correction, and only slightly worse after bias correction.

This pattern remained clear upon applying bias correction, where the AFF+GS and FSL+FS pipelines yielded statistically significantly lower MAE than those without the affine registration. 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. Furthermore, the results are in line with the observation by Cole et al. (2017), who found a substantially reduced between-scanner reliability for a model trained on minimally preprocessed T1w images.

Regardless of the T1w preprocessing differences on the new unseen dataset, the increase in MAE for Models 1 and 4 was 1.5 and 1 year, respectively, even before applying bias correction. This was comparable to or lower than the magnitude of the observed increase in most related literature (Jonsson et al., 2019; Dufumier et al., 2021). Despite the small difference in MAE values, the dataset dependent systematic bias should be mitigated by applying bias correction on a subset of new unseen data, which seems to generally reduce the MAE and also the interquartile range of observed age prediction errors. Note that the order of application of bias correction and ensembling is important, since the reverse order yielded substantially worse performance (results not shown).

5.4. Note on reproducibility

The standardized dataset included multi-site train, validation and test T1w scans of 2504 healthy subjects in the span of 18 to 95 years, and test sets with new site longitudinal T1w scans ($N_{subj} = 2815$). All T1w MRI scans used in the study were obtained from public data sources[7] and were subject to a strict visual quality check to eliminate poor quality scans or scans with failed T1w preprocessing.

In order to enable full reproducibility of the results of this study the lists of included subject IDs and the exact dataset split assignments as used in this study are provided in the Supplementary materials, while 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 are disclosed at the public GitHub repository `https://github.com/AralRalud/BrainAgePreprocessing`.

6. Conclusion

In this paper we studied the effect of preprocessing procedure of T1w MR images 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 statistically significant increase in MAE, up to 0.7 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 statistically 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 MR image 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 new unseen datasets. Regardless of the model or the T1w preprocessing used, some form of bias 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.

[7] Some public data sources may require online registration to gain access to the T1w MRI scans. The UKB dataset is available for a fee.

Acknowledgments

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

- **Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).** 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.
- **Cambridge Centre for Ageing and Neuroscience (CamCAN).** CamCAN funding was provided by the UK Biotechnology and Biological Sciences Research Council (grant number BB/H008217/1), together with support from the UK Medical Research Council and University of Cambridge, UK.
- **OASIS Longitudinal.** Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.
- **ABIDE I.** Primary support for the work by Adriana Di Martino was provided by the (NIMH K23MH087770) and the Leon Levy Foundation.

  Primary support for the work by Michael P. Milham and the INDI team was provided by gifts from Joseph P. Healy and the Stavros Niarchos Foundation to the Child Mind Institute, as well as by an NIMH award to MPM (NIMH R03MH096321).

A. Dataset and model details

A.1. Dataset details

Aim: Train, Validation, Test (Multi-site T1w scans)

Dataset	$\mathbf{N_{scans}}$	**Age span**	$\mu_{\mathbf{age}} \pm \mathbf{sd_{age}}$
ABIDE I[8]	161	18.0 – 48.0	25.7 ± 6.4
ADNI[9]	248	60.0 – 90.0	76.2 ± 5.1
CamCAN (Shafto et al., 2014; Taylor et al., 2017)[10]	624	18.0 – 88.0	54.2 ± 18.4
CC-359 (Souza et al., 2018)[11]	349	29.0 – 80.0	53.5 ± 7.8
FCON 1000[12]	572	18.0 – 85.0	45.3 ± 18.9
IXI[13]	472	20.1 – 86.2	49.0 ± 16.2
OASIS-2 (Marcus et al., 2010)[14]	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**	$\mu_{\mathbf{age}} \pm \mathbf{sd_{age}}$
UK Biobank (Miller et al., 2016)	2815	47.1 – 80.4	63.2 ± 7.2

[1] Data available at: http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html
[2] Data available at: http://adni.loni.usc.edu/
[3] Data available at: https://camcan-archive.mrc-cbu.cam.ac.uk/dataaccess/
[4] Data available at: https://sites.google.com/view/calgary-campinas-dataset/download
[5] Data available at: http://fcon_1000.projects.nitrc.org/indi/enhanced/neurodata.html
[6] Data available at: https://brain-development.org/ixi-dataset/
[7] Data available at: https://www.oasis-brains.org/

Table 4: Age statistics, i.e. span, mean age ($\mu_{\mathbf{age}}$) and associated standard deviation ($\mathbf{sd_{age}}$) in *years*, per dataset of included T1w subject scans in train, test and validation datasets (*top*) and in the new unseen site and test-retest datasets (*bottom*).

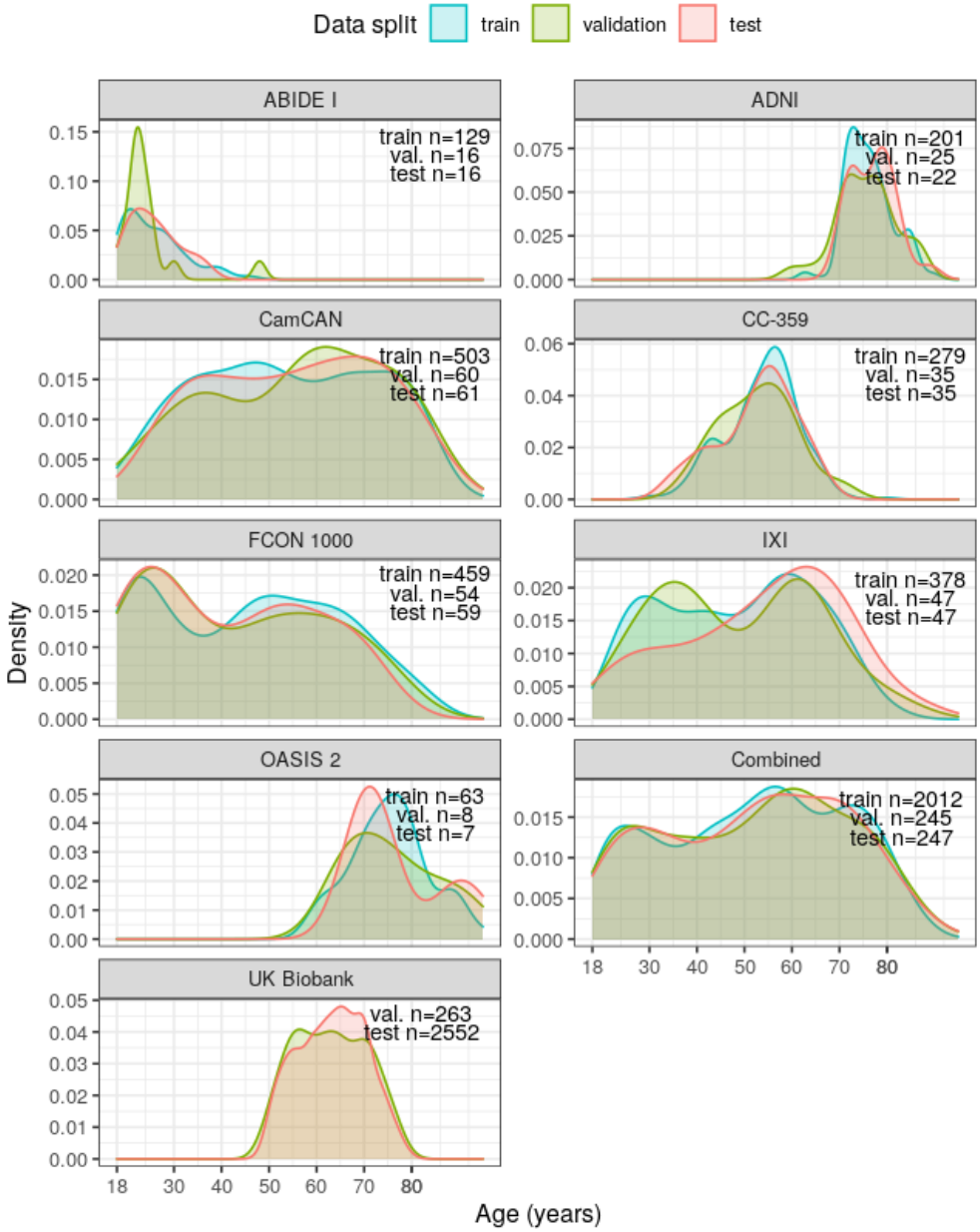

Figure 8: Density of age distribution per each and combined multi-site dataset, depicted for train, test and validation set splits.

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 Figure 9 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 LR values; this was also the case for Models 2 and 3 with the LR values set as proposed in original papers. In general, we observed that training

	Model 1		Model 2	
	Proposed	**Implemented**	**Proposed**	**Implemented**
Input size	$182 \times 218 \times 182$	$157 \times 189 \times 170$	$157 \times 189 \times 15$	
***Batch size**	28	16	16	32
***Loss function**	L1		MSE	L1
***Learning rate (LR)**	1×10^{-2}	1×10^{-4}	1×10^{-4}	1×10^{-3}
LR decay	3%		1×10^{-4}	
Weight decay	5×10^{-5}		1×10^{-3}	
Momentum	0.9		0.9	
Parameters	$\approx$ 900 000		$\approx$ 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	$95 \times 79 \times 78$		$160 \times 192 \times 160$	$157 \times 189 \times 170$
***Batch size**	16	8	8	9
***Loss function**	MSE	L1	KLD	
***Learning rate (LR)**	5×10^{-5}		1×10^{-2}	
LR decay	1×10^{-4}		$\times$0.3 every 30 epochs	
Weight decay	5×10^{-4}		1×10^{-3}	
Momentum	0.9		0.9	
Parameters	$\approx$ 900 000		$\approx$ 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]

Table 5: Proposed hyperparameter values in original literature and the values chosen herein. Only the hyperparameters marked with * were reevaluated. The resulting model accuracy is reported as MAE in years.

with L1 loss was most stable and produced overall lower MAE values, compared to the use of MSE and KLD 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.

A.3. 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 minutes, while the more complex AFF+GS took 4-7 minutes per image. The Fs+FSL pipeline was most time consuming, taking > 15 minutes 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.

B. Linear Mixed Effect Model results

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

ANOVA	F value		**LMEM**	Estimate	Std. Err.	2.5 %	97.5 %
			Intercept	2.971	0.185	2.608	3.334
Model	46.921	***	Model 2	0.561	0.179	0.211	0.912
			Model 3	0.167	0.179	-0.183	0.518
			Model 4	0.199	0.179	-0.152	0.549
Preproc.	6.212	***	Fs+FSL	0.111	0.179	-0.240	0.461
			RIG	0.207	0.179	-0.144	0.557
			RIG+GSC	0.161	0.179	-0.190	0.511
Model:Preproc	1.572		Model 2:Fs+FSL	0.614	0.253	0.118	1.109
			Model 3:Fs+FSL	0.170	0.253	-0.325	0.666
			Model 4:Fs+FSL	0.138	0.253	-0.358	0.634
			Model 2:RIG	0.583	0.253	0.087	1.078
			Model 3:RIG	0.102	0.253	-0.393	0.598
			Model 4:RIG	-0.211	0.253	-0.707	0.284
			Model 2:RIG+GSC	0.449	0.253	-0.047	0.944
			Model 3:RIG+GSC	0.090	0.253	-0.405	0.586
			Model 4:RIG+GSC	-0.178	0.253	-0.674	0.317
			Random effects	Variance	SD		
			Subject ID (Intercept)	4.506	2.123		
			Residual	3.962	1.990		

Table 7: Results of ANOVA and LMEM with absolute error as response variable, and 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 was not statistically significant.

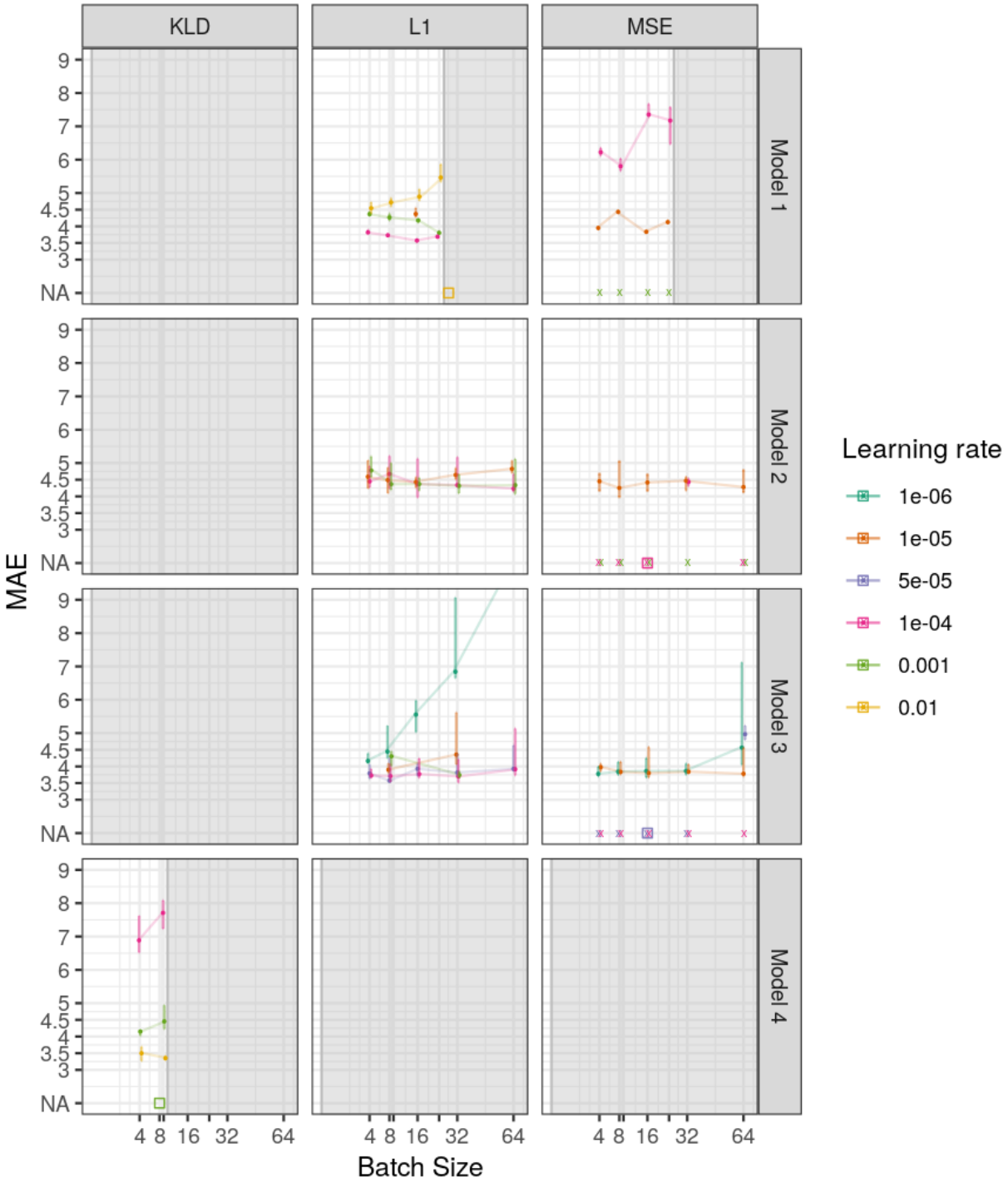

Figure 9: Median, minimal and maximal MAE value of 10 last training epochs for each hyperparameter setting. The hyperparameter values proposed in original research of four models are marked with *square*, the ones resulting in training divergence are *marked as NA and with a cross*. Hyperparameter space for large batch sizes was inaccessible due to hardware limitations and is *grayed out*.

ANOVA	F value		**LMEM**	Estimate	Std. Err.	2.5 %	97.5 %
			Intercept	3.859	0.060	3.740	3.977
Model	406.1692	***	Model 2	0.5663	0.068	0.434	0.699
			Model 3	-0.01959	0.068	-0.152	0.113
			Model 4	-0.136	0.068	-0.269	-0.004
Preproc.	155.3276	***	Fs+FSL	-0.206	0.068	-0.339	-0.074
			RIG	0.600	0.068	0.468	0.733
			RIG+GS	0.692	0.068	0.559	0.824
BC	3103.8325	***	BC	-0.753	0.068	-0.886	-0.621
Model:Preproc.	12.6120	***	Model 2:Fs+FSL	0.745	0.096	0.557	0.932
			Model 3:Fs+FSL	0.167	0.096	-0.020	0.355
			Model 4:Fs+FSL	0.721	0.096	0.534	0.909
			Model 2:RIG	0.088	0.096	-0.099	0.276
			Model 3:RIG	0.339	0.096	0.151	0.526
			Model 4:RIG	0.239	0.096	0.051	0.426
			Model 2:RIG+GS	-0.177	0.096	-0.365	0.010
			Model 3:RIG+GS	0.063	0.096	-0.125	0.250
			Model 4:RIG+GS	-0.265	0.096	-0.452	-0.077
Model:BC	0.1008		Model 2:BC	0.082	0.096	-0.105	0.270
			Model 3:BC	0.117	0.096	-0.070	0.305
			Model 4:BC	0.161	0.096	-0.027	0.348
Preproc.:BC	74.9882	***	Fs+FSL:BC	0.377	0.096	0.189	0.564
			RIG:BC	-0.524	0.096	-0.711	-0.336
			RIG+GS:BC	-0.659	0.096	-0.846	-0.471
Model:Preproc.:BC	16.0805	***	Model 2:Fs+FSL:BC	-0.875	0.135	-1.140	-0.610
			Model 3:Fs+FSL:BC	-0.206	0.135	-0.471	0.060
			Model 4:Fs+FSL:BC	-0.641	0.135	-0.907	-0.376
			Model 2:RIG:BC	0.274	0.135	0.009	0.539
			Model 3:RIG:BC	-0.227	0.135	-0.493	0.038
			Model 4:RIG:BC	-0.185	0.135	-0.450	0.080
			Model 2:RIG+GS:BC	0.297	0.135	0.032	0.562
			Model 3:RIG+GS:BC	0.047	0.135	-0.218	0.313
			Model 4:RIG+GS:BC	0.271	0.135	0.006	0.537
			Random effects	Variance	SD		
			Subject ID (Intercept)	3.400	1.844		
			Residual	5.843	2.417		

Table 8: Results of ANOVA and LMEM tests on UK Biobank dataset preprocessed with the same preprocessing procedure as the training dataset with absolute error as response variable, and model architecture, BC and preprocessing procedure as fixed factor: $Abs\ Error = Model + Preproc. + BC + Model * Preproc + Model * BC + Preproc * BC + Model * Preproc * BC + (1|ID)$.

ANOVA	F value		**LMEM**	Estimate	Std. Err.	2.5 %	97.5 %
			Intercept	3.105	0.052	3.003	3.208
Model	396.225	***	Model 2	0.649	0.048	0.554	0.743
			Model 3	0.098	0.048	0.003	0.192
			Model 4	0.024	0.048	-0.070	0.119
Preproc.	26.891	***	Fs+FSL	0.171	0.048	0.076	0.265
			RIG	0.076	0.048	-0.018	0.171
			RIG+GS	0.033	0.048	-0.061	0.128
Model:Preproc.	9.191	***	Model 2:Fs+FSL	-0.130	0.068	-0.264	0.004
			Model 3:Fs+FSL	-0.038	0.068	-0.172	0.096
			Model 4:Fs+FSL	0.080	0.068	-0.054	0.214
			Model 2:RIG	0.362	0.068	0.228	0.496
			Model 3:RIG	0.111	0.068	-0.023	0.245
			Model 4:RIG	0.054	0.068	-0.080	0.188
			Model 2:RIG+GS	0.120	0.068	-0.014	0.254
			Model 3:RIG+GS	0.110	0.068	-0.024	0.244
			Model 4:RIG+GS	0.007	0.068	-0.127	0.141
			Random effects	Variance	SD		
			Subject ID (Intercept)	3.967	1.992		
			Residual	2.973	1.724		

Table 9: Results of ANOVA and LMEM tests on UK Biobank dataset preprocessed with the same preprocessing procedure as the training multi-site dataset. The results are reduced to bias corrected results. LMEM model is defined as: $Abs\ Error = Model + Preproc. + Model * Preproc + (1|ID)$.

ANOVA	F value		**LMEM**	Estimate	Std. Err.	2.5 %	97.5 %
			Intercept	8.881	0.077	8.730	9.033
BC	38848.42	***	UKB	-5.178	0.091	-5.357	-5.000
Model	2658.49	***	Model 2	3.103	0.091	2.925	3.282
			Model 3	-0.625	0.091	-0.804	-0.447
			Model 4	-2.495	0.091	-2.673	-2.316
Preproc.	6626.95	***	RIG+GS	1.521	0.091	1.342	1.700
			AFF+GS	-4.172	0.091	-4.351	-3.994
			Fs+FSL	-4.317	0.091	-4.496	-4.138
Model:Preproc.	215.78	***	Model 2:RIG+GS	3.683	0.129	3.430	3.936
			Model 3:RIG+GS	3.335	0.129	3.082	3.588
			Model 4:RIG+GS	2.897	0.129	2.644	3.149
			Model 2:AFF+GS	0.757	0.129	0.504	1.010
			Model 3:AFF+GS	3.252	0.129	2.999	3.505
			Model 4:AFF+GS	2.071	0.129	1.818	2.323
			Model 2:Fs+FSL	-1.688	0.129	-1.941	-1.435
			Model 3:Fs+FSL	0.854	0.129	0.602	1.107
			Model 4:Fs+FSL	2.295	0.129	2.042	2.548
Model:BC	1237.10	***	Model 2:UKB	-2.117	0.129	-2.370	-1.865
			Model 3:UKB	0.517	0.129	0.264	0.770
			Model 4:UKB	2.395	0.129	2.143	2.648
Preproc.:BC	5265.70	***	RIG+GS:UKB	-1.586	0.129	-1.839	-1.334
			AFF+GS:UKB	3.732	0.129	3.479	3.984
			Fs+FSL:UKB	3.963	0.129	3.710	4.216
Model:Preproc.:BC	140.78	***	Model 2:RIG+GS:UKB	-3.543	0.182	-3.901	-3.186
			Model 3:RIG+GS:UKB	-3.156	0.182	-3.513	-2.798
			Model 4:RIG+GS:UKB	-2.748	0.182	-3.105	-2.390
			Model 2:AFF+GS:UKB	-1.023	0.182	-1.381	-0.666
			Model 3:AFF+GS:UKB	-2.962	0.182	-3.320	-2.605
			Model 4:AFF+GS:UKB	-1.801	0.182	-2.159	-1.444
			Model 2:Fs+FSL:UKB	1.329	0.182	0.971	1.686
			Model 3:Fs+FSL:UKB	-0.737	0.182	-1.094	-0.379
			Model 4:Fs+FSL:UKB	-2.029	0.182	-2.386	-1.671
			Random effects	Variance	SD		
			Subject ID (Intercept)	4.596	2.144		
			Residual	1.615	3.275		

Table 10: Results of ANOVA and LMEM tests on UK Biobank dataset preprocessed with new preprocessing procedure with absolute error as response variable, and model architecture, BC and preprocessing procedure as fixed factor: $Abs\ Error = BC + Model + Preprocessing + BC * Model + BC * Preprocessing + Model * Preprocessing + Model * Preprocessing * BC + (1|ID)$.

ANOVA	F value		**LMEM**	Estimate	Std. Err.	2.5 %	97.5 %
			Intercept	3.262	0.055	3.155	3.370
Model	589.846	***	Model 2	0.720	0.048	0.625	0.815
			Model 3	0.182	0.048	0.087	0.277
			Model 4	0.170	0.048	0.076	0.265
Preproc.	169.577	***	Fs+FSL	0.087	0.048	-0.008	0.182
			RIG	0.440	0.048	0.346	0.535
			RIG+GS	0.375	0.048	0.281	0.470
Model:Preproc.	20.479	***	Model 2:Fs+FSL	-0.093	0.068	-0.227	0.041
			Model 3:Fs+FSL	-0.173	0.068	-0.307	-0.039
			Model 4:Fs+FSL	-0.003	0.068	-0.137	0.131
			Model 2:RIG	0.266	0.068	0.132	0.400
			Model 3:RIG	-0.290	0.068	-0.424	-0.156
			Model 4:RIG	-0.270	0.068	-0.403	-0.136
			Model 2:RIG+GS	0.406	0.068	0.272	0.540
			Model 3:RIG+GS	-0.111	0.068	-0.245	0.023
			Model 4:RIG+GS	-0.121	0.068	-0.254	0.013
			Random effects	Variance	SD		
			Subject ID (Intercept)	4.743	2.178		
			Residual	2.980	1.726		

Table 11: Results of ANOVA and LMEM tests on UK Biobank dataset preprocessed with new preprocessing. The results are reduced to bias corrected results. LMEM model is defined as: $Abs\ Error = Model + Preproc. + Model * Preproc + (1|ID)$.

CRediT authorship contribution statement

Lara Dular: Conceptualization and evaluation protocol of this study, Data cleanup, Implementation of methods and experiments, Collection and analysis of results, Wrote and revised the manuscript. **Franjo Pernuš:** Wrote and revised the manuscript. **Žiga Špiclin:** Conceptualization and evaluation protocol of this study, Data collection, Wrote and revised the manuscript.

References

K. Franke, C. Gaser, Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease, GeroPsych 25 (2012) 235–245.

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, Frontiers in Neurology 10 (2019).

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, Annals of Neurology 88 (2020) 93–105.

H. G. Schnack, N. E. van Haren, M. Nieuwenhuis, H. E. Hulshoff Pol, W. Cahn, R. S. Kahn, Accelerated Brain Aging in Schizophrenia: A Longitudinal Pattern Recognition Study, AJP 173 (2016) 607–616.

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

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, Clinical Infectious Diseases 73 (2021) 1813–1821.

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 (2017) 1349–1357.

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

D. M. Hedderich, A. Menegaux, B. Schmitz-Koep, R. Nuttall, J. Zimmermann, S. C. Schneider, J. G. Bäuml, M. Daamen, H. Boecker, M. Wilke, C. Zimmer, D. Wolke, P. Bartmann, C. Sorg, C. Gaser, Increased Brain Age Gap Estimate (BrainAGE) in Young Adults After Premature Birth, Front. Aging Neurosci. 13 (2021).

L. Baecker, R. Garcia-Dias, S. Vieira, C. Scarpazza, A. Mechelli, Machine learning for brain age prediction: Introduction to methods and clinical applications, eBioMedicine 72 (2021).

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, volume 11583, International Society for Optics and Photonics, SPIE, 2020, p. 1158303. URL: https://doi.org/10.1117/12.2579630. doi:10.1117/12.2579630.

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

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. URL: https://arxiv.org/abs/2106.01132. doi:10.48550/ARXIV.2106.01132.

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, Neurobiology of Aging 91 (2020) 15–25.

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, 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, Predicting the age of the brain with minimally processed t1-weighted mri data, 2022. URL: https://www.medrxiv.org/content/early/2022/09/09/2022.09.06.22279594. doi:10.1101/2022.09.06.22279594.

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.

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. doi:10.1109/ISBI.2019.8759392.

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. doi:10.1109/ISBI.2017.7950650.

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.

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 Transactions on Medical Imaging 40 (2021) 3400–3412.

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. URL: https://arxiv.org/abs/2103.11695. doi:10.48550/ARXIV.2103.11695.

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

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 (2020) 5014–5027.

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 (2021).

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: Reports 2 (2022) 100137.

M. Tanveer, M. A. 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, 2022. URL: https://arxiv.org/abs/2212.03868. doi:10.48550/ARXIV.2212.03868.

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, Nat Commun 10 (2019) 5409.

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.

M. Modat, D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, S. Ourselin, Global image registration using a symmetric block-matching approach, Journal of Medical Imaging 1 (2014) 1 – 6.

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 (2010) 192–203.

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 (2010) 1310–1320.

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

D. L. Collins, P. Neelin, T. Peters, A. C. Evans, Automatic 3d intersubject registration of mr volumetric data in standardized talairach space, Journal of Computer Assisted Tomography 18 (1994) 192–205.

Laboratory for Computational Neuroimaging, Athinoula A. Martinos Center for Biomedical Imaging., Freesurferwiki: recon-all, 2022. URL: https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all.

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 (2002) 825–841.

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.

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.

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

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, PNAS 116 (2019) 22341–22346.

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, Neurobiology of Aging 56 (2017) 41–49.

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.

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 (2021).

G. Levakov, G. Rosenthal, I. Shelef, T. R. Raviv, G. Avidan, From a deep learning model back to the brain—Identifying regional predictors and their relation to aging, Human Brain Mapping 41 (2020) 3235–3252.

C.-Y. Kuo, T.-M. Tai, P.-L. Lee, C.-W. Tseng, C.-Y. Chen, L.-K. Chen, C.-K. Lee, K.-H. Chou, S. See, C.-P. Lin, Improving Individual Brain Age Prediction Using an Ensemble Deep Learning Framework, Front Psychiatry 12 (2021).

L. Dular, Ž. Špiclin, Improving Across Dataset Brain Age Predictions Using Transfer Learning, in: I. Rekik, E. Adeli, S. H. Park, J. Schnabel (Eds.), Predictive Intelligence in Medicine, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 243–254. doi:10.1007/978-3-030-87602-9_23.

J. Jönemo, M. U. Akbar, R. Kämpe, J. P. Hamilton, A. Eklund, Efficient brain age prediction from 3d mri volumes using 2d projections, 2022. URL: https://arxiv.org/abs/2211.05762. doi:10.48550/ARXIV.2211.05762.

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, Healthy Ageing and Longevity, Springer International Publishing, Cham, 2019, pp. 293–328. URL: https://doi.org/10.1007/978-3-030-24970-0_19. doi:10.1007/978-3-030-24970-0_19.

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).

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 (2017) 262–269.

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.

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, Journal of Cognitive Neuroscience 22 (2010) 2677–2684.

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 (2016) 1523–1536.