NeuroImage: Clinical 37 (2023) 103319

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Age-level bias correction in brain age prediction

Biao Zhang [a], Shuqin Zhang [a,*], Jianfeng Feng [b], Shihua Zhang [c,*]

[a] School of Mathematical Sciences, Fudan University, Shanghai 200433, China
[b] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China
[c] NCMIS, CEMS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Keywords:
Age prediction
Bias correction
Human brain
Machine learning
MRI

ABSTRACT

The predicted age difference (PAD) between an individual's predicted brain age and chronological age has been commonly viewed as a meaningful phenotype relating to aging and brain diseases. However, the systematic bias appears in the PAD achieved using machine learning methods. Recent studies have designed diverse bias correction methods to eliminate it for further downstream studies. Strikingly, here we demonstrate that bias still exists in the PAD of samples with the same age even after kind of correction. Therefore, current PAD may not be taken as a reliable phenotype and more investigations are needed to solve this fundamental defect. To this end, we propose an age-level bias correction method and demonstrate its efficacy in numerical experiments.

1. Introduction

Brain aging often accompanies cognitive decline and dementia, and even neurological diseases (Cole et al., 2018) such as Alzheimer's disease (Abbott, 2011), schizophrenia (Koutsouleris et al., 2014), and Parkinson's disease (Reeve et al., 2014). Thus abnormal brain aging is usually considered an important indicator of the occurrence of such diseases. As an individual's brain age is often different from his or her chronological age, computational prediction based on brain magnetic resonance imaging (MRI) data has been a common way of estimating brain age (Cole et al., 2018). Machine learning methods including feature extraction-based shallow learning (Franke et al., 2010; Wang et al., 2014; Kondo et al., 2015; Cole et al., 2015; Liem et al., 2017) and end-to-end deep learning methods (Huang et al., 2017; Cole et al., 2017; Jónsson et al., 2019; Peng et al., 2021; Cheng et al., 2021) have been applied for this task.

The predicted age difference (PAD) between the predicted brain age and chronological age (Jónsson et al., 2019), sometimes referred to as brain age delta (Cole et al., 2017; Smith et al., 2019), has been proposed to characterize how an individual deviates from a healthy brain aging trajectory (Fig. 1). Several studies have shown that high positive PAD correlates with neurological degeneration and the development of diseases, such as lower fluid intelligence and higher mortality (Cole et al., 2018), traumatic brain injuries (Cole et al., 2015), cognitive impairments (Franke et al., 2012; Liem et al., 2017) and schizophrenia (Koutsouleris et al., 2014; Schnack et al., 2016), while negative PAD is related to a healthy lifestyle. Thus, PAD has been viewed as an important phenotype relating to brain diseases (e.g., Alzheimer's disease, brain injury), physical activity and even genome sequence variants (Cole et al., 2017; Jónsson et al., 2019; Kaufmann et al., 2019) (Fig. 1).

However, there exists a systematical bias in the predicted age for subjects of all ages, indicating an over-prediction of the age for relatively younger individuals and an under-prediction for elderly individuals (de Lange and Cole, 2020; Smith et al., 2019). For general nonlinear prediction methods, the real cause of the bias is still obscure. Le et al. (2018) have shown that this bias is inevitable for regression, rather than a property limited to age prediction. It has been defined as 'regression dilution', which is attributed to the non-Gaussian distribution of the chronological age (MacMahon et al., 1990; Fuller, 2009; Smith et al., 2019). However, when the age prediction method is linear, e.g., ordinary linear regression (OLS), for chronological age Y, the bias is generated due to the fact that the predicted age $\widehat{Y}$ and the PAD $= \widehat{Y} - Y$ are orthogonal, which forces PAD and Y to have an angle between 0 and 90 degrees (Habeck et al., 2017). For models that do not account for significant variance in age Y, PAD and age Y will be more obviously correlated. This explanation of bias in linear situation reflects the cause of PAD in nonlinear cases. Since PAD is supposed to be an informative index that tells scientists or clinicians how a person compares his/her own age to peers in terms of brain health, and ideally provides predictive utility independent of chronological age, the correlation between uncorrected PAD and age undoubtedly weakens the rationality of PAD as a biomarker or phenotype.

* Corresponding authors.
E-mail addresses: zhangb20@fudan.edu.cn (B. Zhang), zhangs@fudan.edu.cn (S. Zhang), Jianfeng64@gmail.com (J. Feng), zsh@amss.ac.cn (S. Zhang).

https://doi.org/10.1016/j.nicl.2023.103319
Received 3 July 2022; Received in revised form 28 November 2022; Accepted 4 January 2023
Available online 7 January 2023

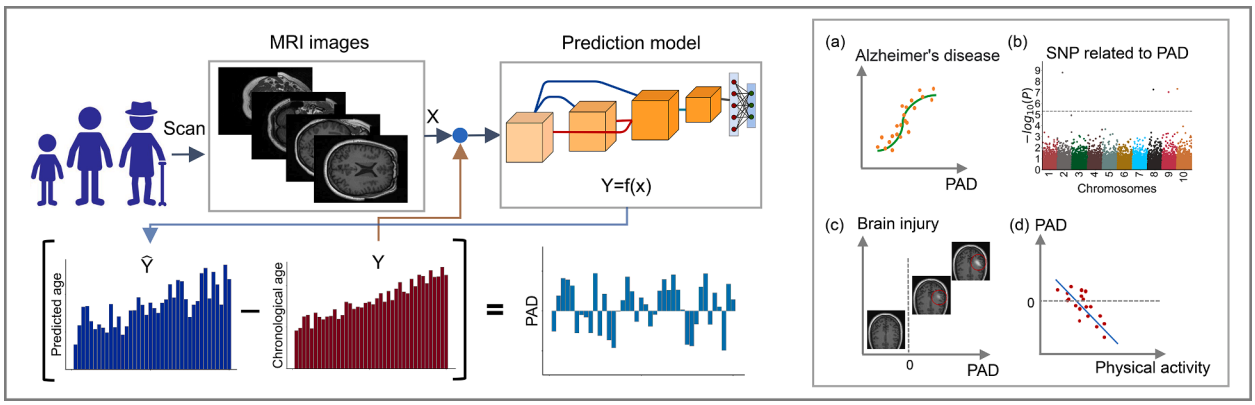

Fig. 1. Illustration of the computation and applications of PAD. Left: The predicted age is obtained by a prediction model trained on the brain data (e.g., structural MRI image) X and the chronological age Y of the samples. PAD is defined as the difference between the predicted age $\hat{Y}$ and the chronological age Y. Right: PAD has been considered an important phenotype relating to brain diseases (e.g., Alzheimer's disease, brain injury), physical activity, and even genome sequence variants.

To eliminate the bias existing in the predicted brain age, several bias correction methods have been developed (Beheshti et al., 2019; de Lange et al., 2019; de Lange and Cole, 2020; Liang et al., 2019; Smith et al., 2019; Treder et al., 2021). Bias correction is usually executed as an additional step after the prediction of brain age. Linear correction methods are commonly used, while high-order correction methods such as quadratic correction show similar results to the linear ones (Smith et al., 2019). There are mainly two linear correction methods, i.e., the Cole's method and Beheshti's method (de Lange et al., 2020). And such linear bias correction methods can be easily adapted to nonlinear ones, e.g., by replacing the linear regression in these methods with the quadratic regression (Smith et al., 2019). Although some recent methods add bias correction constraints for the regression model such as LASSO during model training (Treder et al., 2021), some studies claimed that this kind of methods essentially adjusts the degree of linear bias correction after training and provides a balance between Mean Absolute Error (MAE) and PAD bias. Although those bias correction methods have been adopted to correct the bias in the PAD of all samples (sample-level bias), which gives the mean of PAD over all samples close to zero, however, in this paper, we show that after such bias corrections, the bias appears significantly in the PAD of samples with the same age (age-level bias). This phenomenon exists for various datasets, age prediction methods, and sample-level bias correction methods. The existence of age-level bias weakens the reliability of results in previous research related to PAD. Therefore, we propose an age-level correction method and verify its efficacy for different settings. To the best of our knowledge, this is the first time to consider age-level bias. Furthermore, via doing OLS regression between non-imaging indexes in UK Biobank and two variables: chronological age and corrected PAD, we show that the age-level corrected PAD is a potentially reliable phenotype.

2. Methods

2.1. Datasets and preprocessing

We used three brain MRI datasets including UK Biobank (Miller et al., 2016), OASIS (LaMontagne et al., 2019), and ABIDE (Craddock et al., 2013). UK Biobank is a large-scale biomedical database, which contains multi-modal brain image data of people in UK. We followed the data processing pipeline of the UK Biobank in Alfaro-Almagro et al. (2018). OASIS (v3) is a dataset containing T1w MRI data of more than 1000 participants that were collected across 30 years. Participants include 609 cognitively normal adults and 489 individuals at various stages of cognitive decline ranging in age from 42 to 97. We used 3388 T1 structural MRI images from 1098 subjects. We directly used the processed data by the OASIS team. ABIDE is an MRI dataset containing functional and structural brain imaging data collected from multiple laboratories for studying the neural bases of autism. We used 1099 T1 structural MRI images. We directly used the processed data from the ABIDE I provided by the ABIDE website. Since the edges of MRI images are often empty, the 3D MRI images in all three datasets were cropped to the proper sizes. Table 1 shows the basic information of all these three datasets. The 41 non-imaging indexes in UK Biobank we considered are listed in Table 3 of the reference Smith et al. (2019).

In some studies, the bias correction model is fitted using the training data other than the test one, on which PAD is computed, while most other methods assume the chronological ages are known and the bias correction models are fitted using the test data directly. However, we showed that the bias correction results were almost identical no matter using an independent validation dataset or not for fitting parameters for both the Cole's and Beheshti's methods (Supplementary Fig. 8). Therefore, in the following, we adopt the setting without an independent validation dataset for bias correction step. For age prediction accuracy measured by MAE, we achieved the minimum MAE of 2.55 years by ResNet with Kullback–Leibler divergence loss, which is comparable to the minimum MAE achieved by other studies on dataset UK Biobank (Peng et al., 2021).

2.2. Age prediction methods

2.2.1. Loss functions for deep neural networks

Kullback–Leibler divergence (KL) Peng et al. (2021) transformed the chronological age to a probability vector with a fixed length and trained the neural network by minimizing the KL divergence between the probability vectors of the chronological age and the predicted brain age. Suppose the length of the age probability vector is K, for sample i, the chronological age Y_i can be represented as the mean of the age vector, that is

Table 1
Summary of three brain age estimation datasets

Dataset	Sample Size	Age Range	Age Statistics (Mean ± STD)	Cropped Size	Training	Test	Validation
UK Biobank	9880	[38, 86]	62.02 ± 7.48	(160, 192, 160)	7979	1482	419
OASIS	3388	[42, 97]	66.92 ± 9.70	(160, 196, 224)	2575	678	135
ABIDE	1099	[6, 65]	17.07 ± 8.03	(224, 224, 160)	836	220	43

$$Y_i = \sum_{k=1}^{K} p_{ik} \cdot a_{ik}, \tag{1}$$

where a_{ik} is the age vector for sample i in a dataset (e.g., the age vector of UK Biobank is $[38, 86]$). p_{ik} is generated by a Gaussian distribution with variance equal to 1. For sample i, the predicted age $\widehat{Y}_i$ can also be represented as the mean of the age vector with an estimated probability vector, that is

$$\widehat{Y}_i = \sum_{k=1}^{K} \widehat{p}_{ik} \cdot a_{ik}, \tag{2}$$

where $\widehat{p}_{ik}$ is the kth element of the probability vector for sample i estimated by neural network. The total KL loss is

$$\mathrm{KL}(\{Y_i\}, \{\widehat{Y}_i\}) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}(P_i \,|\, \widehat{P}_i) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}\big(P_i \,\big|\, f_p(X_i)\big), \tag{3}$$

where N is the number of samples and f_p represents the neural network, which outputs a probability vector.

Mean Square Error (MSE) It is defined as

$$\mathrm{MSE}(\{Y_i\}, \{\widehat{Y}_i\}) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - f(X_i))^2, \tag{4}$$

where f represents the neural network that outputs the predicted age directly.

Cross-Entropy loss (CE) When cross-entropy loss is used, neural networks also output a softmax probability vector. Besides, the chronological age is rounded to integers and the regression problem is transformed into a classification problem. The cross-entropy loss is defined as

$$\mathrm{CE}\left(\left\{Y_i\right\}, \left\{\widehat{Y}_i\right\}\right) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} (-Y_{ik} \log(\widehat{p}_{ik})), \tag{5}$$

where $\widehat{p}_{ik}$ is the kth element of the estimated softmax probability vector, and $(Y_{i0}, Y_{i1}, \ldots, Y_{iK})$ is the one-hot vector formulation of Y_i.

2.2.2. Deep learning methods

The deep learning methods were all trained on NVIDIA Tesla V100 GPU. The data processing procedures are identical. We used learning rate decay for all the models and chose the optimal epoch number by evaluating PAD on the validation dataset. The detailed hyper-parameters of all the deep learning methods for various datasets and loss functions are summarized in Supplementary Table 1.

3D ResNet We implemented the ResNet (He et al., 2016) in Peng et al. (2021) for 3D images and modified the last layer slightly for different datasets or loss functions. ResNet is a special convolutional neural network with short connections between layers. The convolutional filters mostly are $3 \times 3 \times 3$ and batch normalization layer (Ioffe and Szegedy, 2015) is used in almost every layer. Considering the problem complexity, we chose ResNet with 34 layers (ResNet-34). Besides, for different loss functions, the hyper-parameters were optimized on the validation dataset. To be specific, the final 3D average pooling is set to $(3, 6, 5)$ for UK Biobank, $(5, 6, 7)$ for OASIS, and $(7, 7, 5)$ for ABIDE, respectively. To ensure the convergence of neural networks, we added a nonlinear ReLU activation layer into the linear layer when the loss is MSE and CE. On UK Biobank, 3D ResNet-34 achieves an MAE of 2.55 years with KL loss, 2.77 years with MSE loss, and 2.81 years with CE loss. On OASIS, 3D ResNet-34 achieves an MAE of 2.23 years with KL loss. On ABIDE, 3D ResNet-34 achieves an MAE of 3.38 years with KL loss.

SFCN Simple Fully Convolutional Network (SFCN) (Peng et al., 2021) defeats other methods on age prediction tasks in Predictive Analytic Challenge (PAC) 2019. The model consists of seven blocks and each of the first five blocks contains a $3 \times 3 \times 3$ 3D convolutional layer, a batch normalization layer, a max pooling layer and a ReLU activation layer. The sixth block contains a $1 \times 1 \times 1$ 3D convolutional layer, a batch normalization layer and a ReLU activation layer. The seventh block contains an average pooling layer, a dropout layer (Srivastava et al., 2014) (50% dropout rate), a fully connected layer and a softmax output layer. We used the default parameters apart from adjusting the batch size to make SFCN converge on the UK Biobank dataset. In Peng et al. (2021), when SFCN and 3D ResNet share the same training parameters, they achieve comparative performance in the training set. However, after hyper-parameters adjustment, 3D ResNet performs better on all three data sets we used. In our experiment, SFCN achieves an MAE of 3.17 years with KL loss.

3D MSDNet Mixed-scale dense convolutional neural network (MSDNet) (Pelt and Sethian, 2018) has been shown to be effective on large image segmentation with significantly fewer parameters and training samples. Since a single MRI image has millions of voxels, and most MRI datasets are composed of an insufficient quantity of images, we adapted the architecture of MSDNet to the age prediction problem. In the original MSDNet, feature maps of each layer are connected to the other layers, and the shape of the feature map keeps the same across layers. As shown in Supplementary Fig. 1, we added multiple blocks into the 3D MSDNet, and each block has the same structure as the original 3D MSDNet. Then, between any two blocks, there is a max-pooling operation scaling down the 3D MRI images. At the last layer, the 3D feature maps are flattened to be input into a fully-connected neural network. 3D MSDNet achieves an MAE of 3.88 years with KL loss.

2.2.3. Statistical learning methods

For statistical learning methods, we first applied the $2 \times 2 \times 2$ max-pooling and flattening on the MRI image data in UK Biobank. Then we used PCA to extract 1000 features with maximal variance. We further trained Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996), Support Vector Regression (SVR) (Smola and Schölkopf, 2004), and XGBoost (Chen and Guestrin, 2016) using package `scikit-learn`, with the training dataset of UK Biobank and tested their performance on the test dataset. LASSO is a regression analysis method that performs both variable selection and regularization. SVR is developed for function estimation based on Support Vector Machines (SVM) (Cortes and Vapnik, 1995). XGBoost is an extended end-to-end algorithm of gradient boosting tree and it is used widely on many machine learning challenges. The three methods achieve 3.94, 4.01, and 3.93 of MAE with KL loss. And the hyper-parameters were optimized on the validation dataset.

2.3. Bias correction methods

2.3.1. Sample-level bias correction

There are mainly two linear correction methods, i.e., the Cole's method (Cole et al., 2018; Peng et al., 2021; Smith et al., 2019) and Beheshti's method (de Lange et al., 2020). Specifically, let $Y, \widehat{Y}, \widehat{Y}_c$ represent the chronological age, predicted age, and predicted age with correction, respectively. Let $\mathrm{PAD} = \widehat{Y} - Y$ and $\mathrm{PAD}_c = \widehat{Y}_c - Y$ denote the PAD and the corrected one, respectively. The Cole's method first regresses $\widehat{Y}$ on Y to estimate the linear relations between the predicted age and chronological age using

$$\widehat{Y} = \alpha \times Y + \beta, \tag{6}$$

where α and β represent the slope and intercept used to correct the predicted age, respectively. Then PAD is corrected by

$$\mathrm{PAD}_c = \widehat{Y}_c - Y = \frac{\widehat{Y} - \beta}{\alpha} - Y.$$

The Beheshti's method first fits the relationship between PAD and the chronological age as

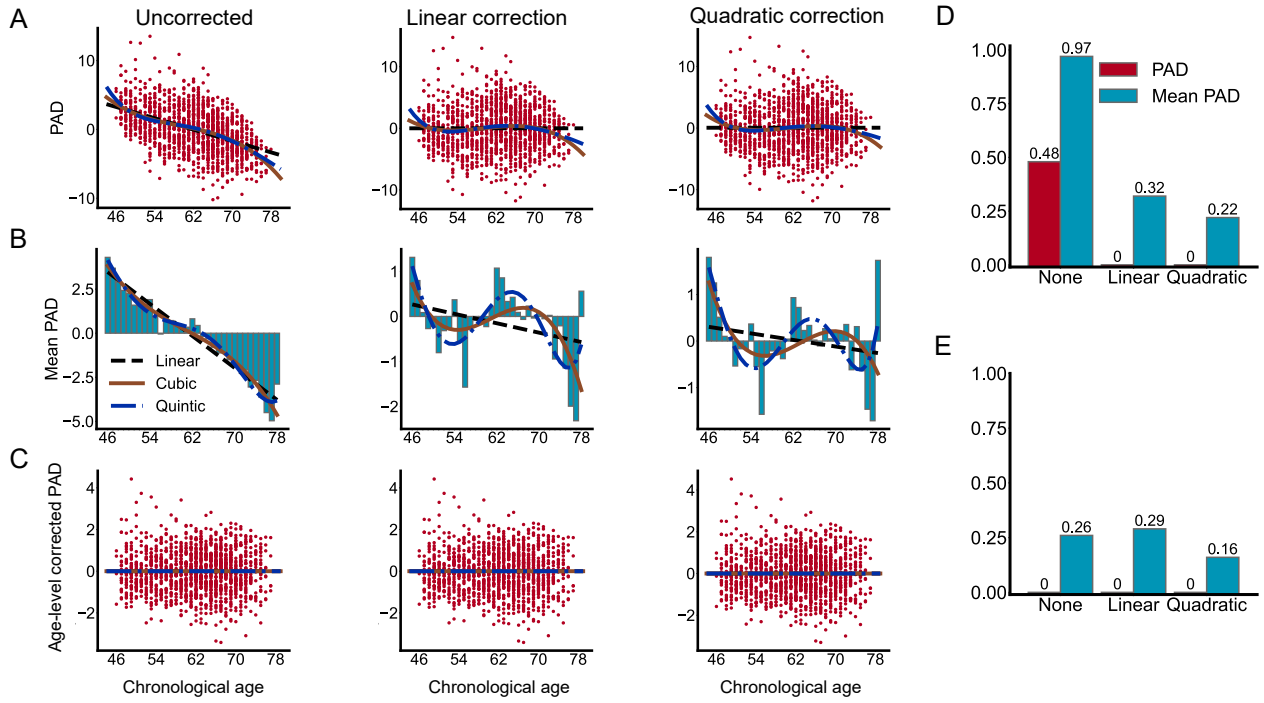

Fig. 2. Illustration of significant discrepancy between PAD, mean PAD of the same age, and age-level corrected PAD. The prediction model is ResNet with the KL divergence loss trained on the UK Biobank dataset. 'Uncorrected','Linear correction', and 'Quadratic correction' mean PAD is uncorrected or corrected using Cole's method with linear or quadratic correction, respectively. **A.** The scatter plots of PAD and the corrected PAD. **B.** The bar plots of the mean of PAD and the corrected PAD over samples of the same age. The trend curves are fitted with the linear, cubic, and quintic polynomials, respectively. **C.** The scatter plots of age-level corrected PAD. Age-level corrected PAD displays almost no bias patterns whether PAD is first corrected using Cole's method or not. **D.** Comparison of the Pearson correlations between PAD, mean PAD of the same age and chronological age without or with bias correction using Cole's method, respectively. **E.** Comparison of the Pearson correlations between PAD, mean PAD of the same age and chronological age, respectively. PAD is age-level corrected after bias correction using Cole's method or not.

$$\mathrm{PAD} = \alpha \times Y + \beta,$$

and the PAD is corrected by

$$\mathrm{PAD}_c = \widehat{Y}_c - Y = \widehat{Y} - [(\alpha + 1) \times Y + \beta].$$

Besides, de Lange et al. (2019) adopts an equivalent one to the Beheshti's method after deriving α and β using the same method as that in Eq. (6), and PAD is corrected by

$$\mathrm{PAD}_c = \widehat{Y}_c - Y = \widehat{Y} - (\alpha \times Y + \beta).$$

These bias correction methods have no significant differences except that the data corrected by the Cole's method inevitably contains higher variance as the predicted age is divided by the slope value α for each subject, while the Beheshti's method reduces the variance and results in a lower MAE as it includes the chronological age in the correction.

2.3.2. Age-level bias correction

To eliminate the bias that still exists after applying the well-known bias correction methods, we propose a straightforward age-level bias correction method. It corrects the bias via eliminating the bias curve corresponding to the mean PAD of samples at each age after the sample-level bias correction. For samples of age a, let μ_a, σ_a denote the mean, standard deviation of PAD over samples of age a, respectively, we can correct the PAD of sample i at the age level via

$$\mathrm{PAD}_i^{\mathrm{ac}} = \left(\mathrm{PAD}_i - \mu_a\right)/\sigma_a, \tag{7}$$

where $\mathrm{PAD}_i^{\mathrm{ac}}$ denotes the age-level corrected PAD of the same age a. This kind of correction can be executed after the Cole's method or Beheshti's method. The bias could be eliminated with this straightforward correction, since the mean of $\mathrm{PAD}_i^{\mathrm{ac}}$ of the same age a is zero:

$$\mathbb{E}_a\left[\mathrm{PAD}_{\mathrm{i}}^{\mathrm{ac}}\right] = \mathbb{E}_a[(\mathrm{PAD}_i - \mu_a)/\sigma_a] = (\mathbb{E}_a[\mathrm{PAD}_i] - \mu_a)/\sigma_a = 0.$$

2.4. Data and code availability

The UK Biobank dataset is accessible upon applications via the website: https://www.ukbiobank.ac.uk/. OASIS can be downloaded from the website: https://www.oasis-brains.org/. ABIDE can be downloaded from the website: https://fcon_1000.projects.nitrc.org/indi/abide/. The code of this paper is deposited on GitHub: https://github.com/saulgoodenough/pad_bias_correction.

3. Results

In this section, we first demonstrate that age-level bias still exists after applying the current bias correction methods in PAD across multiple data sets, several up-to-date machine learning methods, and different loss functions. We then show that with the proposed age-level bias correction method, the correlation between PAD and chronological age is greatly weakened.

3.1. Bias still exists in the PAD of samples with the same age

The bias correction methods including the quadratic ones have been adopted to correct the bias in the PAD of all samples (sample-level bias), which gives the mean of PAD over all samples close to zero. However, here we demonstrate that after such bias corrections, the bias appears significantly in the PAD of samples with the same age (age-level bias). The PAD correction results on the UK biobank dataset using the Cole's method fitted with the linear, cubic, and quintic curves, respectively are shown in Fig. 2A-B. We can clearly observe that a systematic age-level bias pattern appears, though the sample-level bias declines close to

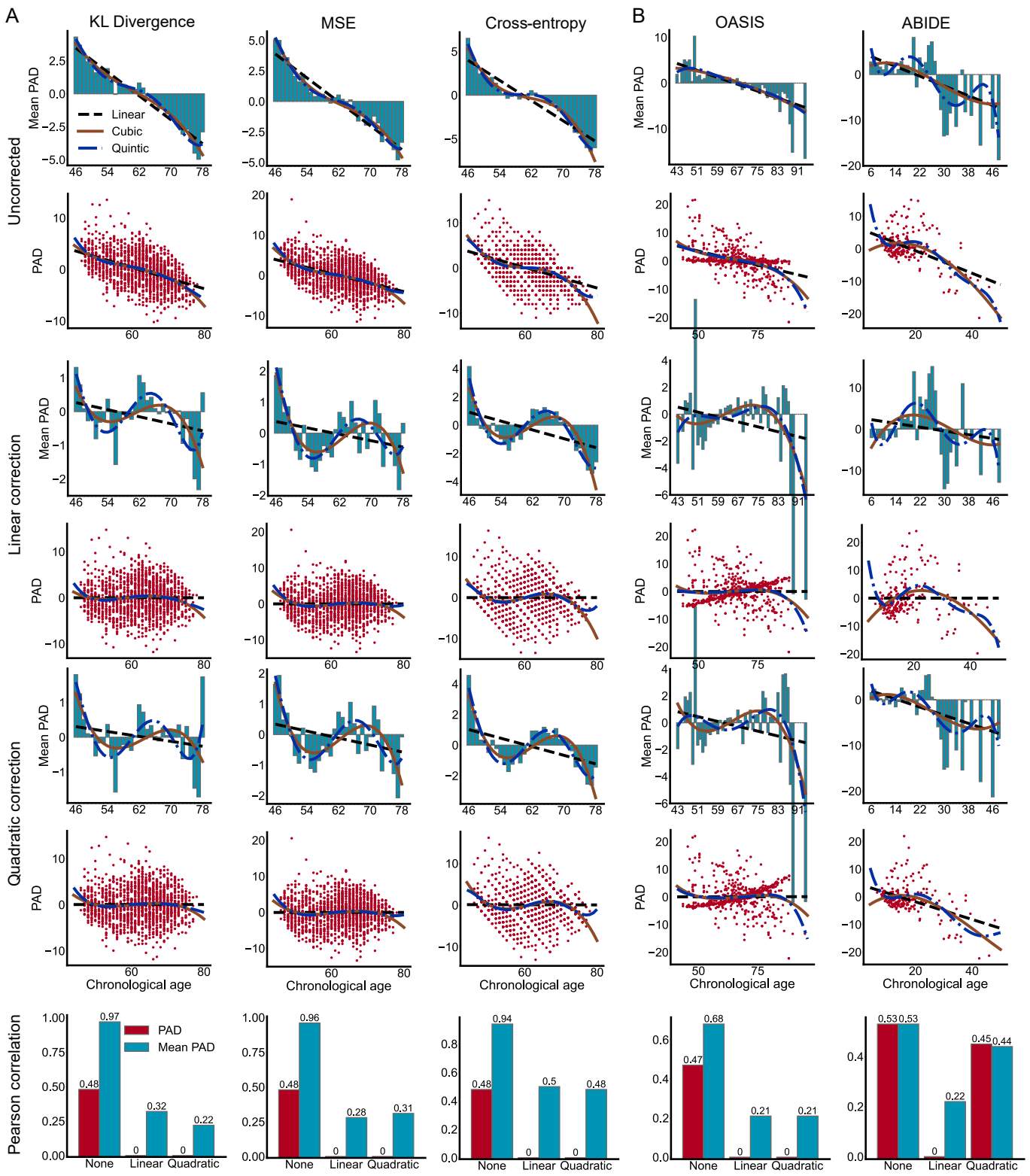

Fig. 3. Illustration of significant discrepancy between PAD and mean PAD of the same age after bias correction with the Cole's method. **A-B.** For different loss functions (KL, MSE, and CE), and datasets (UK Biobank, OASIS, and ABIDE) with 3D ResNet-34, the fitted linear, cubic and quintic curves are quite significant in the bar plot of the mean PAD though all are close to a straight line for the PAD. In bar plots at the bottom, Pearson correlations between PAD and the chronological age decline close to zero while those between mean PAD and chronological age are still very high. These show linear and quadratic bias corrections do not correct the tendency in mean PAD, although they correct bias in the PAD of samples.

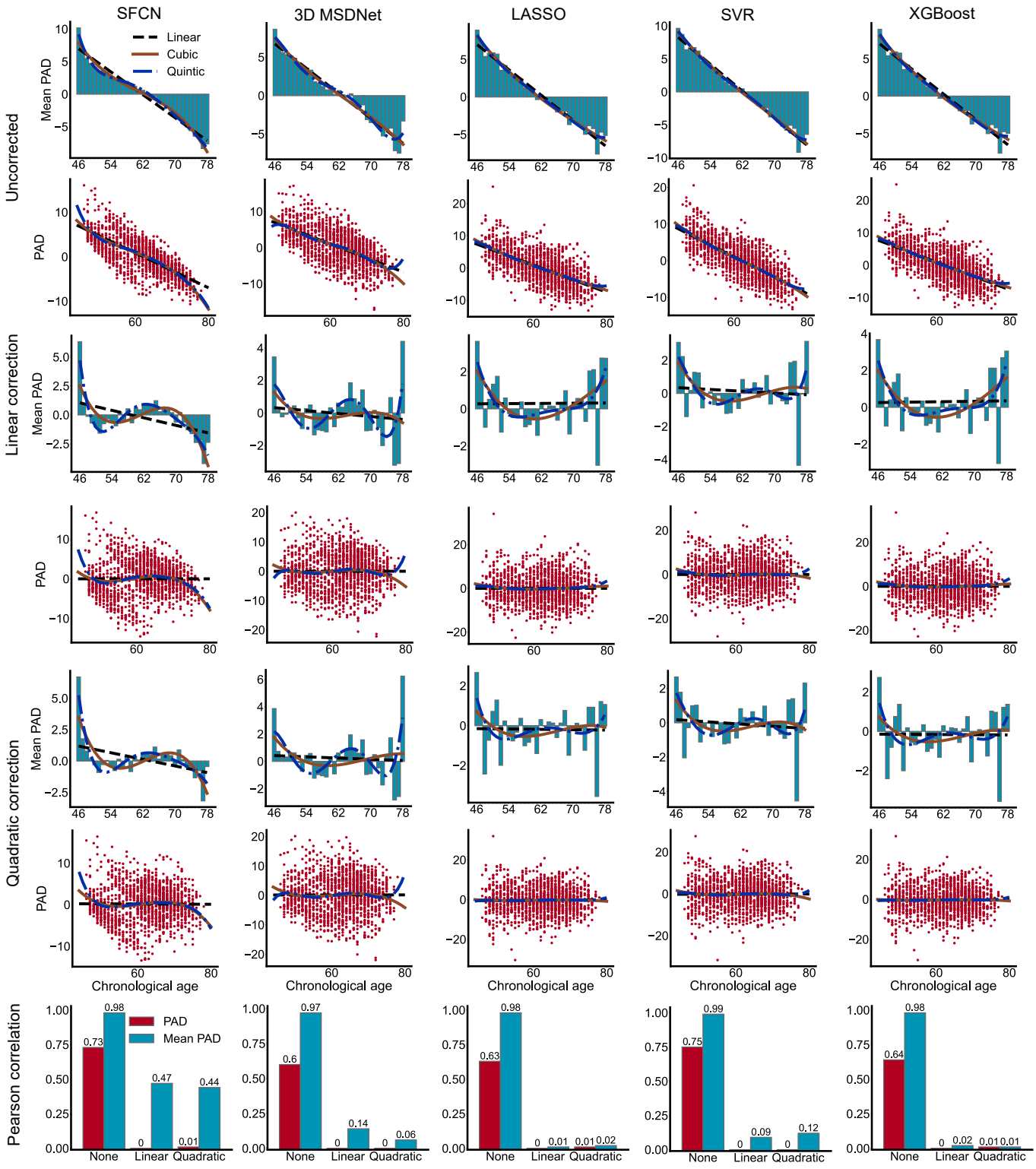

Fig. 4. Illustration of significant discrepancy between PAD and mean PAD of the same age after bias corrections with the Cole's method for different methods (SFCN with KL loss, 3D MSDNet with KL loss, LASSO, SVR and XGBoost). Pearson correlations between PAD and the chronological age decline close to zero while those between mean PAD of the same age and the chronological age are still high for deep learning methods.

zero after linear or quadratic corrections. This phenomenon exists across diverse datasets, methods including deep learning and statistical approaches, and loss functions (Fig. 3 and Fig. 4). The situation of the Beheshti's method is quite similar (Supplementary Fig. 3).

In addition, whether bias exists or not in PAD can be evaluated quantitatively by the correlation between the corrected PAD and the chronological age called PAD correlation (PADC), which is also referred to as age delta correlation (ADC) in Treder et al. (2021). To this end, we calculated the Pearson correlation coefficients (PCC) between the PAD over all samples and their chronological age, and between the mean PAD of samples with the same age and the chronological age, respectively. The age-level PADC is still relatively high after both the linear and

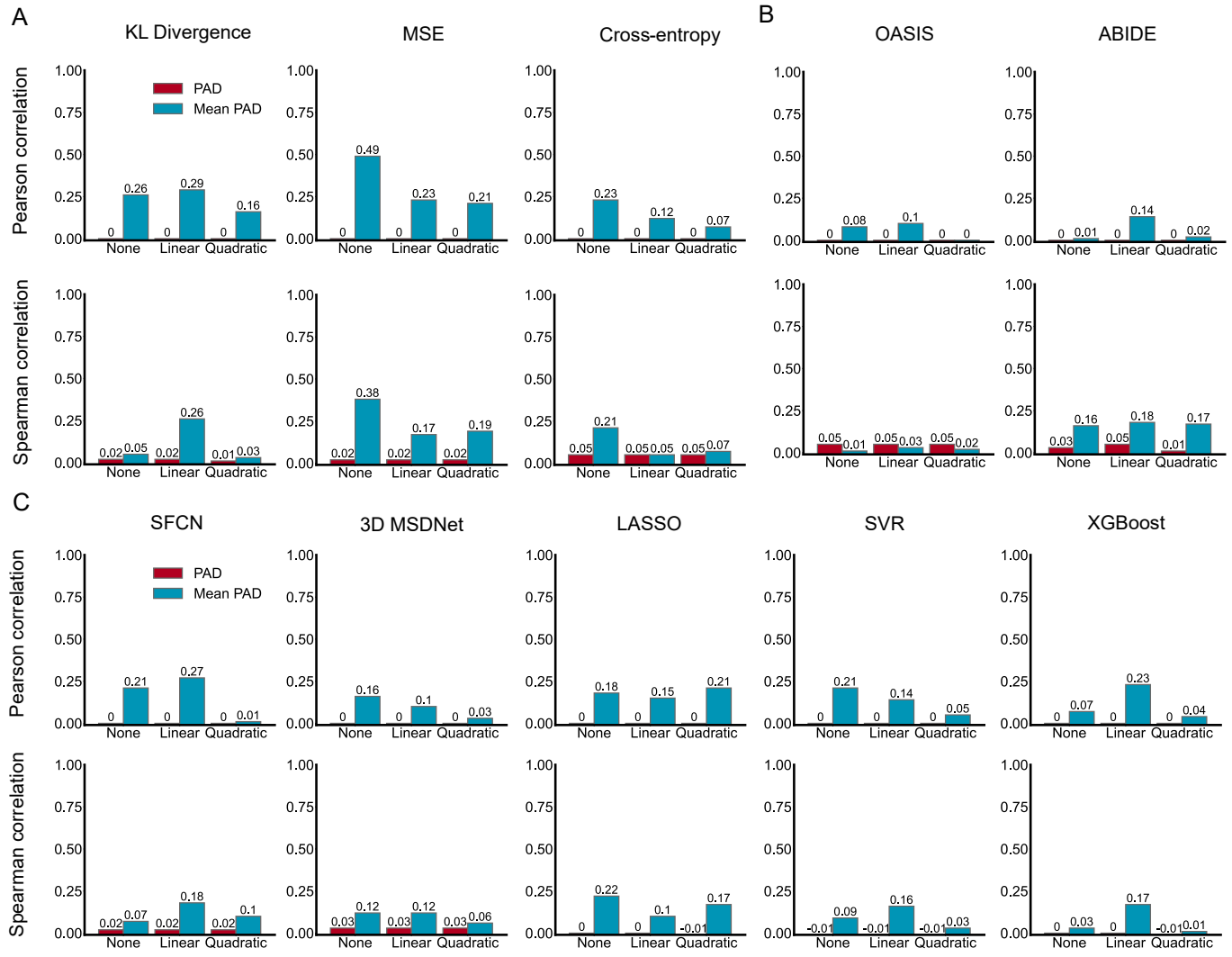

Fig. 5. Age-level bias correction results. Pearson and Spearman correlations between PAD, mean PAD, and chronological age by first using Cole's method followed by age-level correction method for loss functions of deep neural networks (KL loss, MSE loss, and cross-entropy loss) using 3D ResNet-34, different datasets (UK Biobank, OASIS and ABIDE) using 3D ResNet-34 with KL loss and methods (SFCN, 3D MSDNet, LASSO, SVR and XGBoost). Compared to the results in Fig. 3 and Fig. 4, age-level bias correction gives much smaller correlations between PAD, mean PAD, and chronological age.

quadratic corrections using Cole's method though the sample-level PADC almost declines to zero (Fig. 2D). We also used the Spearman rank correlation coefficients (SRCC) to further confirm our findings (Supplementary Fig. 1). This situation is quite similar across diverse correction methods, datasets, loss functions, and prediction methods (Figs. 3, 4, Supplementary Figs. 2–4). These results imply that previous bias correction methods are not sufficient to eliminate the intrinsic correlation between PAD and chronological age.

3.2. Age-level bias correction

Our investigation reveals that the age-level PAD bias correction is quite different from that of the sample level. The existing bias correction methods mainly focus on sample-level bias, while overlook the age-level bias. This intrinsic problem could bring false conclusions in downstream applications. For example, those genome-wide association studies (GWAS) of PAD may yield misleading sequence variants (Cole et al., 2017; Jónsson et al., 2019). Thus, PAD may not be a reliable phenotype correlating with neurological diseases as shown in many studies (Cole et al., 2018). How to correct this special bias requires further exploration. Combining that the sample-level bias is explained via regression dilution resulted from random measurement error (Hutcheon et al., 2010), the age-level bias is presumably caused by random measurement error and variation in the population.

Scatter plots of age-level corrected PAD and chronological age display almost no bias patterns no matter whether sample-level correction is conducted (Fig. 2C, Supplementary Figs. 5–7). Compared to the usual corrected PAD, the age-level corrected PAD gives much fewer correlations between PAD, mean PAD and the chronological age measured by both the Pearson and Spearman correlations in most cases (Fig. 5 and Supplementary Figs. 6–7). An exception is that for LASSO, SVR and XGboost, the mean PAD is already close to zero with only linear or quadratic bias corrections, and then PAD corrected by combinations of age-level and linear or quadratic correlates slightly stronger with chronological age. This should be caused by the linearity and underfitting as the performance of these three models is significantly worse than the other methods as shown in the method section. This is also worth further studying.

Furthermore, we conducted experiments to test if the age-level corrected PAD, which has no linear associations with chronological age, is an independent phenotype reflecting the human health state. We did OLS regression between clinical/cognitive indexes and two variables, i.e., chronological age and PAD corrected by six bias correction methods. We used 41 non-imaging indexes in UK Biobank, which are reported as

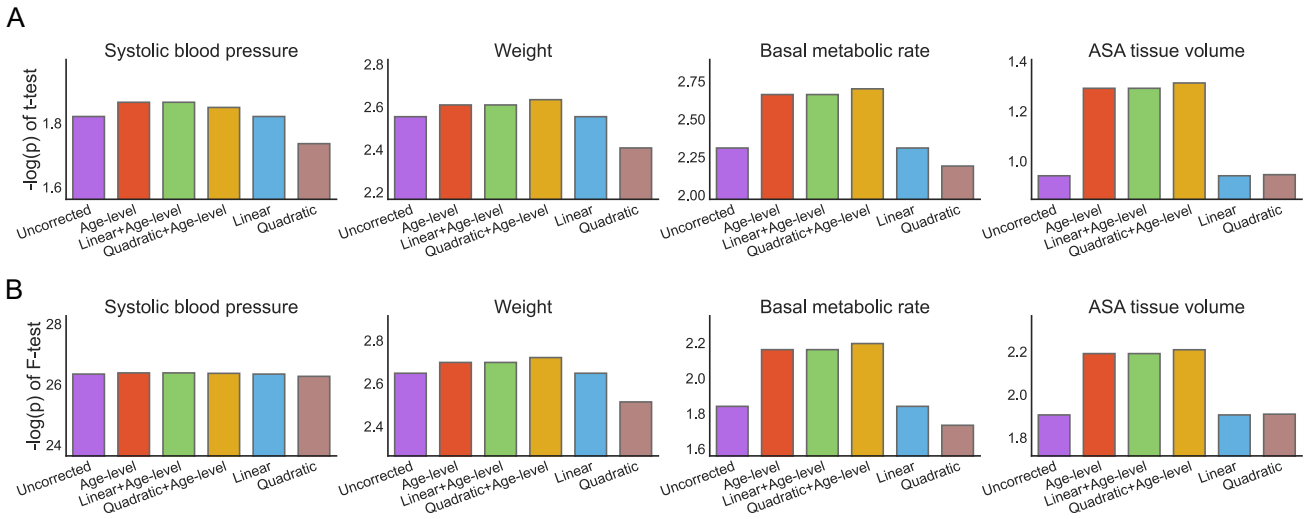

Fig. 6. Comparison of linear and quadratic correction of Cole's method, age-level correction, and their combinations. Statistical significance of *t*-test (**A**) of the corrected PAD coefficient, and *F*-test (**B**) of the linear regression between four non-imaging indexes and two variables, i.e., chronological age and PAD corrected by various methods. The age prediction model is ResNet-34 with KL loss, and the dataset is UK Biobank.

correlating with PAD the most (Smith et al., 2019). In the six methods, 'Uncorrected' is the uncorrected PAD, 'Age-level' represents the age-level corrected PAD, 'Linear' and 'Quadratic' represent linearly and quadratically corrected PAD, respectively. 'Linear + Age-level' and 'Quadratic + Age-level' are the combination of age-level and both linearly and quadratically corrected PAD. The age prediction method is ResNet-34 with KL loss.

To test the significance of the regression, we did *F*-test and computed the coefficient of the determinant (R^2) of each regression model and endpoint, respectively. To check the linear relationship between the response variable and the corrected PAD, we also implemented a *t*-test for the regression coefficient corresponding to the corrected PAD for the six correction methods. The results are presented as bar plots and box plots in Fig. 6 and Supplementary Fig. 10 for Cole's method and Supplementary Fig. 11 for Beheshti's method. As illustrated in Fig. 6B, Supplementary Fig. 10B, and Supplementary Fig. 11, *F*-test shows strong statistical significance ($p - value \leqslant 10^{-2}$) for almost all the regressions. The R^2 values are mostly lower than 0.2 (Supplementary Fig. 10–11), which is consistent with the results in the previous study (Smith et al., 2019), and this indicates that there should be other variables in the regression. Most importantly, for some clinical or cognitive indexes, such as systolic blood pressure, weight, Basal metabolic rate, Abdominal subcutaneous adipose (ASA) tissue volume, etc., the *t*-test *p*-values for the corrected PAD coefficients in the OLS regression increase significantly after age-level corrections (Fig. 6A, Supplementary Fig. 10–11). This implies that the age-level corrected PAD correlates more strongly with those clinical or cognitive indexes. Hence, age-level corrected PAD should be a better phenotype linearly independent of chronological age and reflects the human health state.

The above results show that the straightforward age-level bias correction method performs well in mitigating the age-level bias, though several issues need to be investigated further, for example, the accurate estimation of the mean and variance of PAD requires a considerable number of samples. Besides, developing regression methods with elaborately-designed regularization terms is also a potential way to solve it.

4. Conclusion

In this paper, by doing comprehensive experiments on various brain MRI data sets, we reveal that age-level bias still exists in age prediction models after applying the updated bias correction methods, and suggest an age-level correction strategy. The age-level bias has not been found and properly corrected. As a consequence, many previous studies on brain age prediction are probably not reliable, and applying age-level bias correction to those works is more likely to give quite different results. Promising future work is to make those comparisons. For example, it is meaningful to investigate how the yielded sequence variants (Cole et al., 2017; Jónsson et al., 2019) from GWAS associated with PAD, sample-level corrected PAD, and age-level corrected PAD differ. We mainly focus on the analysis of bias in brain age prediction problems in this paper. However, age prediction is not constrained by brain MRI data. Preceding research combines multi-modal brain data (Niu et al., 2020; Rokicki et al., 2021) including MRI, resting-state functional MRI, diffusion tensor imaging, etc. In this work, the proposed age-level bias correction method is solely for linear bias. Whether there exists some nonlinear bias requires more investigations, and the corresponding age-level correction method also needs further exploration. How age-level bias fluctuates along the data type is also a promising open question. In addition, for general problems similar to brain age prediction, a natural question is whether there exists a bias similar to age-level bias as we described. The sphere can be extended to general regression problems and deserves more attention in machine learning. How to build theoretical explanations of age-level bias is an essential issue. Our simple correction method and its efficacy in experiments suggest that uncertainty analysis and measurement is a potential approach. Sample-level correction method (Hahn et al., 2022) based on uncertainty analysis demonstrates its superiority. Besides, although some sample-level correction methods train prediction models by adding the correlation between the corrected PAD and the chronological age into the objective function or constraints, how to execute age-level bias correction during the model training is unknown. For deep learning, lack of interpretability further increases the problem's hardness. In summary, further investigations are still necessary for age-level bias correction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my code/data in the manuscript.

Acknowledgments

This study was conducted under the UK Biobank application number of 19542. Shuqin Zhang was partially supported by Science and Technology Commission of Shanghai Municipality (Grant No. 20ZR1407700). This work has been supported by the National Key Research and Development Program of China [No. 2019YFA0709501 to Shihua Zhang], and the CAS Project for Young Scientists in Basic Research [No. YSBR-034 to Shihua Zhang].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.nicl.2023.103319.

References

Cole, James H, Ritchie, Stuart J, Bastin, Mark E, Valdés Hernández, M.C., Muñoz Maniega, S., Royle, Natalie, Corley, Janie, Pattie, Alison, Harris, Sarah E, Zhang, Qian, et al., 2018. Brain age predicts mortality. Mol. Psychiatry 23 (5), 1385–1392.

Abbott, Alison, 2011. Dementia: a problem for our age. Nature 475 (7355), S2–S4.

Koutsouleris, Nikolaos, Davatzikos, Christos, Borgwardt, Stefan, Gaser, Christian, Bottlender, Ronald, Frodl, Thomas, Falkai, Peter, Riecher-Rössler, Anita, Möller, Hans-Jürgen, Reiser, Maximilian, et al., 2014. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophrenia Bull. 40 (5), 1140–1153.

Reeve, Amy, Simcox, Eve, Turnbull, Doug, 2014. Ageing and parkinson's disease: why is advancing age the biggest risk factor? Ageing Res. Rev. 14, 19–30.

Franke, Katja, Ziegler, Gabriel, Klöppel, Stefan, Gaser, Christian, Initiative, Alzheimer's Disease Neuroimaging, et al., 2010. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. Neuroimage 50 (3), 883–892.

Wang, Jieqiong, Li, Wenjing, Miao, Wen, Dai, Dai, Hua, Jing, He, Huiguang, 2014. Age estimation using cortical surface pattern combining thickness with curvatures. Med. Biol. Eng. Comput. 52 (4), 331–341.

Kondo, Chihiro, Ito, Koichi, Kai, Wu., Sato, Kazunori, Taki, Yasuyuki, Fukuda, Hiroshi, Aoki, Takafumi, 2015. An age estimation method using brain local features for t1-weighted images. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) IEEE, pp. 666–669.

Cole, James H, Leech, Robert, Sharp, David J, 2015. Alzheimer's Disease Neuroimaging Initiative. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Ann. Neurol. 77 (4), 571–581.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., Riedel-Heller, S., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage 148, 179–188.

Huang, T.W., Chen, H.T., Fujimoto, R., Ito, K., Wu, K., Sato, K., Taki, Y., Fukuda, H., Aoki, T., 2017. Age estimation from brain MRI images using deep learning. IEEE, pp. 849–852.

Cole, James H, Poudel, Rudra PK, Tsagkrasoulis, Dimosthenis, Caan, Matthan WA, Steves, Claire, Spector, Tim D, Montana, Giovanni, 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124.

Jónsson, Benedikt Atli, Gyda Bjornsdottir, T.E., Thorgeirsson, Lotta María, Ellingsen, G Bragi, Walters, DF Gudbjartsson, Stefansson, Hreinn, Stefansson, Kari, Ulfarsson, M. O., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nature Commun. 10 (1), 1–10.

Peng, Han, Gong, Weikang, Beckmann, Christian F, Vedaldi, Andrea, Smith, Stephen M, 2021. Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. 68, 101871.

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

Smith, Stephen M, Vidaurre, Diego, Alfaro-Almagro, Fidel, Nichols, Thomas E, Miller, Karla L, 2019. Estimation of brain age delta from brain imaging. Neuroimage 200, 528–539.

de Lange, Cole, J.H., 2020. Commentary: Correction procedures in brain-age prediction. NeuroImage: Clinical 26.

de Lange, A.M.G., Kaufmann, T., van der Meer, D., Maglanoc, L.A., Alnæs, D., Moberget, T., Douaud, G., Andreassen, O.A., Westlye, L.T., 2019. Population-based neuroimaging reveals traces of childbirth in the maternal brain. Proceedings of the National Academy of Sciences 116 (44), 22341–22346.

Franke, Katja, Luders, Eileen, May, Arne, Wilke, Marko, Gaser, Christian, 2012. Brain maturation: predicting individual brainage in children and adolescents using structural mri. Neuroimage 63 (3), 1305–1312.

Schnack, Hugo G, Van Haren, Neeltje EM, Nieuwenhuis, Mireille, Hulshoff, Hilleke E, Pol, Wiepke Cahn, Kahn, René S, 2016. Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study. Am. J. Psychiatry 173 (6), 607–616.

Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A., Bettella, F., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nature neuroscience 22 (10), 1617–1623.

Le, Trang T, Kuplicki, Rayus T, McKinney, Brett A, Yeh, Hung-Wen, Thompson, Wesley K, Paulus, Martin P, Aupperle, Robin L, Bodurka, Jerzy, Cha, Yoon-Hee, Feinstein, Justin S, et al., 2018. A nonlinear simulation framework supports adjusting for age when analyzing brainage. Front. Aging Neurosci. 10, 317.

MacMahon, S., Peto, R., Collins, R., Godwin, J., Cutler, J., Sorlie, P., Abbott, R., Neaton, J., Dyer, A., Stamler, J., 1990. Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. The Lancet 335 (8692), 765–774.

Fuller, W.A., 2009. Measurement error models. John Wiley & Sons.

Habeck, C., Razlighi, Q., Yunglin Gazes, D., Barulli, Jason Steffener, Stern, Yaakov, 2017. Cognitive reserve and brain maintenance: orthogonal concepts in theory and practice. Cereb. Cortex 27 (8), 3962–3969.

Treder, Matthias S, Shock, Jonathan P, Stein, Dan J, DuPlessis, Stefan, Seedat, Soraya, Tsvetanov, Kamen A, 2021. Correlation constraints for regression models: controlling bias in brain age prediction. Front. Psychiatry 12, 25.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., 2018. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400–424.

Beheshti, Iman, Nugent, Scott, Potvin, Olivier, Duchesne, Simon, 2019. Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. NeuroImage: Clinical 24, 102063.

Liang, Hualou, Zhang, Fengqing, Niu, Xin, 2019. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. Technical report, Wiley Online Library.

Miller, Karla L, Alfaro-Almagro, Fidel, Bangerter, Neal K, Thomas, David L, Yacoub, Essa, Junqian, Xu., Bartsch, Andreas J, Jbabdi, Sa.ad., Sotiropoulos, Stamatios N, Andersson, Jesper LR, et al., 2016. Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature Neurosci. 19 (11), 1523–1536.

LaMontagne, Pamela J, Benzinger, Tammie LS, Morris, John C, Keefe, Sarah, Hornbeck, Russ, Xiong, Chengjie, Grant, Elizabeth, Hassenstab, Jason, Moulder, Krista, Vlassenko, Andrei, et al., 2019. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv.

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M., Yan, C., 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics 7, 27.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Ioffe, S. and Szegedy, C., 2015, June. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, Salakhutdinov, Ruslan, 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Pelt, D.M., Sethian, J.A., 2018. A mixed-scale dense convolutional neural network for image analysis. Proceedings of the National Academy of Sciences 115 (2), 254–259.

Tibshirani, Robert, 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288.

Smola, Alex J, Schölkopf, Bernhard, 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.

Chen, Tianqi, Guestrin, Carlos, 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Cortes, Corinna, Vapnik, Vladimir, 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Hutcheon, J.A., Chiolero, A., Hanley, J.A., 2010. Random measurement error and regression dilution bias. Bmj, p. 340.

Niu, Xin, Zhang, Fengqing, Kounios, John, Liang, Hualou, 2020. Improved prediction of brain age using multimodal neuroimaging data. Human Brain Mapp. 41 (6), 1626–1643.

Rokicki, J., Wolfers, T., Nordhøy, W., Tesli, N., Quintana, D.S., Alnæs, D., Richard, G., de Lange, A.M.G., Lund, M.J., Norbom, L., Agartz, I., 2021. Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders. Human brain mapping 42 (6), 1714–1726.

Hahn, T., Ernsting, J., Winter, N.R., Holstein, V., Leenings, R., Beisemann, M., Fisch, L., Sarink, K., Emden, D., Opel, N., Redlich, R., 2022. An uncertainty-aware, shareable, and transparent neural network architecture for brain-age modeling. Science advances 8 (1), eabg9471.