

# Replication and Refinement of Brain Age Model for adolescent development

Bhaskar Ray<sup>1,2</sup>, Jiayu Chen<sup>1</sup>, Zening Fu<sup>1</sup>, Pranav Suresh<sup>1,2</sup>, Bishal Thapaliya<sup>1,2</sup>, Britny Farahdel<sup>1,2</sup>, Vince D. Calhoun<sup>1,2</sup>, and Jingyu Liu<sup>\*1,2</sup>

<sup>1</sup>Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS) Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, USA

<sup>2</sup>Department of Computer Science, Georgia State University, Atlanta, USA

## Abstract

The discrepancy between chronological age and estimated brain age, known as the brain age gap, may serve as a biomarker to reveal brain development and neuropsychiatric problems. This has motivated many studies focusing on the accurate estimation of brain age using different features and models, of which the generalizability is yet to be tested. Our recent study has demonstrated that conventional machine learning models can achieve high accuracy on brain age prediction during development using only a small set of selected features from multimodal brain imaging data. In the current study, we tested the replicability of various brain age models on the Adolescent Brain Cognitive Development (ABCD) cohort. We proposed a new refined model to improve the robustness of brain age prediction. The direct replication test for existing brain age models derived from the age range of 8-22 years onto the ABCD participants at baseline (9 to 10 years old) and year-two follow-up (11 to 12 years old) indicate that pre-trained models could capture the overall mean age failed precisely estimating brain age variation within a narrow range. The refined model, which combined broad prediction of the pre-trained model and granular information with the narrow age range, achieved the best performance with a mean absolute error of 0.49 and 0.48 years on the baseline and year-two data, respectively. The brain age gap yielded by the refined model showed significant associations with the participants' information processing speed and verbal comprehension ability on baseline data.

**Keywords:** Brain age estimation, Multi-modal data, Modeling, Replication, Structural MRI

## 1 Introduction

As we grow old, our brain develops and changes parallelly throughout our life. Extracting insightful information from neuroimaging data underlying brain development has become a widely studied research area recently [1–3]. An individual's brain age is referred to as biological age, which cannot be measured directly but can be modeled through neuroimaging data. Brain age models are usually trained with brain imaging features of magnetic resonance imaging (MRI) or other types of neuroimaging resources from a healthy or typical development population. The brain age gap is computed by the difference between chronological age and estimated brain age. Recently brain age gap has been utilized as a potential biological indicator for identifying several neurodegenerative disorders such as Alzheimer's and Parkinson's at their early stage [4–6], or for understanding brain development trajectory or abnormal aging process [7–9].

Multiple studies have been conducted using various methods such as voxel-based, region-based, and surface-based techniques to obtain features from different modalities to accurately predict brain age. The early studies of brain age were mainly focused on using traditional machine learning techniques such as regression-based approaches to estimate brain age [10–13]. For instance, the authors of [11]

---

\*Corresponding author, jliu75@gsu.edu

applied a support vector regression (SVR) model with 1,116,006 features derived from multimodal structural imaging maps of 621 subjects from the Philadelphia Neurodevelopmental Cohort (PNC) cohort aged 8-22 years and achieved a mean absolute error (MAE) of 1.22 years on estimated brain age. Lund et al. [13] used a single modality, resting-state fMRI imaging data, to analyze the brain development of adolescents from the PNC cohort and measured the relationship between individuals' cognitive ability and brain age gap. Based on the shrinkage estimation of the regression model, the MAE was 2.43 years with a correlation of 0.60 between the actual age and the predicted age. When the same model was tested on an independent sample of subjects in the same age group (8-22 years) from the healthy brain network cohort, it achieved an MAE of 3.46 years and a correlation value of  $r = 0.58$ .

Deep learning models have recently gained popularity in the brain age research field due to their ability to produce consistent promising results in various imaging modalities [14–19]. Peng et al. [14] developed a simple, fully convolutional neural network by analyzing 14,503 T1 weighted structural MRI (sMRI) images of UK biobank data and achieved an MAE of 2.14 years in brain age estimation as a part of the 2019 predictive analysis challenge for brain age prediction. The authors of [15] constructed a novel convolutional neural network model by adding more filters in the initial phase of the architecture for more accurate brain age estimation (N=1445 T1 weighted sMRI images). These prior brain age studies mostly focused on predicting brain age using either single- or multi-modal structural or functional MRI data. However, the generalizability or robustness of these brain age models has not been comprehensively tested using independent samples.

One primary goal of this study is to test the replicability and generalizability of the pre-trained brain age models built in our previous study [1]. These brain age models were trained with the PNC dataset [20], consisting of 1417 adolescent subjects aged 8-22 years. Our previous study has shown that a traditional machine learning regression model with a small set of selected important features from a multimodal dataset can achieve high brain age prediction accuracy, similar to or better than those from complicated deep learning neural network models. In the current work, we investigated how previously trained brain age models perform on the Adolescent Brain and Cognitive Development (ABCD) [21–23] cohort. The ABCD participants have a narrow age range (9-10 years at baseline and 11-12 years at year-two follow-up). Furthermore, we also developed a refined brain age model that helps improve brain age models' generalizability and prediction accuracy.

The rest of the paper is organized as follows. The materials and methods section describes the cohort, data preprocessing, and feature extraction procedure. The data analyses section presents detailed analytical approaches to evaluate PNC pre-trained brain age models, newly trained ABCD brain age models, and refined brain age models, followed by the association tests between cognitive ability and the brain age gap, and identification of the vital brain regions for brain age prediction. The test results and visualization of important brain regions are presented in the Results section. Finally, we provided the insights and knowledge we learned from our results in the Discussion and Conclusion section.

## 2 MATERIALS AND METHODS

### 2.1 Participants

The ABCD study (<https://abcdstudy.org/>) is a longitudinal study intended to run for 10 years. Over 11 thousand children at 9-10 years were recruited from 21 sites across the USA to collect a wide range of data, including brain imaging data, genetics, cognitive and behavioral assessments, family history and environmental measures, etc. The institutional review board approved protocols were followed to obtain all participating children's consent and their parent's full written informed consent. The current study used the available data from ABCD release 2.0, including baseline and year-two follow-up. Specifically, we have analyzed T1-weighted sMRI images of 11573 children from the ABCD baseline and T1-weighted sMRI images of 2947 children aged 11-12 years from the ABCD year-two follow-up.

## 2.2 Data preprocessing

We preprocessed sMRI images using the Statistical Parametric Mapping 12 (SPM12) software to generate the gray matter density maps. SPM12 jointly segments and spatially normalizes the T1 MRI data using SPM12 default tissue probability maps and generates six types of tissue maps (grey matter, white matter, CSF, bone, soft tissue, and others) in the Montreal Neuroimaging Institute (MNI) space. The gray matter maps were further smoothed by a  $6\text{ mm}^3$  Gaussian kernel and resliced to a voxel size of  $1.5 \times 1.5 \times 1.5\text{ mm}^3$ . Quality control was applied to individual gray matter maps to select those correlated to the group mean gray matter map greater than 0.9, which resulted in 11,382 participants' images in baseline (male: 5,960, female: 5,422; age:  $9.92 \pm 0.63$ ), 2877 in year-two follow-up (male: 1553, female: 1324; age:  $11.93 \pm 0.63$ ). Then we applied the same gray matter mask from our previous PNC study [1], on ABCD data to include the voxels with gray matter density larger than 0.2.

## 2.3 Feature extraction

The same sets of brain features used in our previous PNC brain age model were extracted for the ABCD cohort at baseline and year-two follow-up from preprocessed sMRI data. Specifically, we extracted three sets of brain structural features: 1) gray matter loadings of 100 brain components derived by independent component analysis (ICA) [24], 2) gray matter density of 116 regions of interest (ROI) based on the automatic anatomical labeling (AAL) [25] atlas, and 3) 152 anatomical features extracted by Freesurfer v5.3, software [26]. First, we have projected the same 100 ICA brain components used in our PNC brain age model onto ABCD baseline and two-year follow-up data. In our previous work, we applied the ICA model:  $X_{PNC} = A_{PNC} \times S_{PNC}$ .  $X_{PNC}$  denotes the gray matter data matrix,  $A_{PNC}$  is the loading matrix with each column indicating the expression of each component on individual subjects, and  $S_{PNC}$  is the 100 independent brain sources or component matrix of PNC cohort. Here we applied the projection equation  $A_{ABCD} = X_{ABCD} \times S_{PNC}^{-1}$  onto the ABCD cohort data to get the new loading matrix,  $A_{ABCD}$ , which shows how the previous 100 ICA components derived from PNC data are expressed in the new ABCD cohort. It was conducted separately for ABCD baseline and year-two follow-up data. Therefore, the matrix,  $A_{ABCD}$ , presenting loadings of gray matter components contains the feature vectors used for further brain age estimation. Secondly, we have used the AAL atlas map to extract the mean gray matter density features of 116 brain regions from the ABCD cohort. Thirdly, the ABCD study has provided tabulated data including FreeSurfer-derived imaging anatomical features for both baseline and year-two follow-up [23]. We selected the 152 anatomical features, which include estimated total intracranial volume, cortical thickness, and cortical and subcortical volumes of the left and right hemispheres of the brain based on the Desikan atlas.

# 3 Data Analyses for Brain Age models

## 3.1 PNC pre-trained brain age models

Our previous study was conducted using the data collected from PNC with adolescents aged 8-22 years. The primary goal of our PNC study was to utilize multimodal brain imaging data for brain age prediction and provide a detailed comparison of the brain age models' performance trained on different unimodal or multimodal imaging modalities. Four types of imaging features were analyzed, including the three feature sets from gray matter images focusing on the brain structural features and one from the resting-state functional MRI – the loadings of 100 ICA components from derived functional network connectivity strength. To estimate the brain age, two sets of regression models, support vector regression (SVR, linear kernel) and partial least square regression with and without recursive feature elimination (RFE) were trained and tested using the brain imaging features as the independent predictors and the subject's chronological age as the dependent variable. Details of the model building can be seen in reference [1]. In summary, we found that the SVR models with multimodal data and feature selection produced the best prediction accuracy. The multimodal SVR model trained with only 120 features selected from brain structural feature sets achieved an MAE of 1.25 years. The SVR

model with selected 188 features from all four feature sets obtained an MAE of 1.22 years. Since the inclusion of the features from functional MRI data in the multimodal brain age model only provided limited improvement in the prediction accuracy, we focus on testing the reliability and generalizability of the brain age models only using structural brain features in the current study by excluding the functional data.

In this current research, we tested the performance of the pre-trained SVR brain age models with the ABCD baseline and year-two follow-up data. To comprehensively understand model reliability (e.g., whether individual feature sets have different reliability), we investigated eight brain age models listed in Table 1, named based on the features used. Table 1 summarizes the brain features used for individual models and their accuracy on PNC data reported in the previous study. For simplicity, we will refer to each brain age model as the short name in Table 1. The ABCD subjects have a narrow age range of 9-10 years and 11-12 years for baseline and year-two, respectively. We first evaluated all PNC pre-trained SVR brain age models using ABCD brain structural features at baseline, and then evaluated *FreeSurfer\_152*, *FreeSurfer\_RFE\_110*, *Combined\_368*, and *Combined\_RFE\_120* models using ABCD year-two follow up data as these four models showed reasonable performance for baseline data as explained later. For the same reason, feature sets from these four models were used to build new models explained in the following as a comparison.

Table 1: Description of the Brain age models.

Pre-trained SVR Models	Features	Accuracy on PNC data	
		<i>Train</i> (MAE)	<i>Test</i> (MAE)
ICA_100	Loadings of 100 ICA components from gray matter images	1.69 ± 0.23	1.68
ICA_RFE_81	RFE selected 81 ICA components	1.64 ± 0.25	1.60
ROI_116	Gray matter density of 116 AAL brain regions	1.77 ± 0.18	1.90
ROI_RFE_70	Gray matter density of RFE selected 70 AAL brain regions	1.76 ± 0.19	1.73
FreeSurfer_152	152 brain anatomical features from FreeSurfer	1.77 ± 0.14	1.76
FreeSurfer_RFE_110	RFE selected 110 brain anatomical features from FreeSurfer	1.70 ± 0.30	1.68
Combined_368	368 combined brain structural features: 100 ICA components, gray matter density of 116 AAL regions, 152 brain anatomical features	1.54 ± 0.22	1.32
Combined_RFE_120	RFE selected 120 combined brain structural features: 40 ICA components, mean density of 38 AAL brain regions, 42 anatomical features	1.32 ± 0.13	1.25

### 3.2 ABCD self-trained brain age models

To verify the predictive power of the ABCD data and to put the performance of pre-trained models into perspective, we newly trained four brain age models using FreeSurfer or combined brain features from ABCD baseline data. These four models used the same feature sets as in *FreeSurfer\_152*, *FreeSurfer\_RFE\_110*, *Combined\_368*, and *Combined\_RFE\_120*. We divided the baseline samples into 90% training and 10% testing data. Then on the 90% training data, we applied 10-fold cross-evaluation with the grid-search method to select the best hyper-parameters and best-fitted SVR models. The newly trained ABCD brain age models were tested on the 10% hold-out testing data, the full baseline samples, and the year-two follow-up samples. When directly testing the replication of PNC pre-trained models, we applied the models to the full samples of ABCD baseline and year-two follow-up. To make comparisons fair, we thus focused on the test performance of ABCD self-trained models and later refined brain age models using the full samples of ABCD baseline and year-two follow-up. We have reported the mean of predicted brain age, Pearson correlation coefficient (r-value) between actual age and predicted brain age, and MAE in years of predicted brain age to evaluate the performance of various brain age models.

### 3.3 Refined brain age models

We constructed refined brain age models (see Figure 1) to improve the robustness of existing brain age models. The refined model first leveraged the broader prediction power of PNC pre-trained models by feeding imaging features to the PNC pre-trained models. Then the refined model added a new module to capture the unexplained residuals, which are the differences between the actual age and the predicted age from the broader brain age models. Finally, the refined brain age was obtained by subtracting the estimated residuals from the broader predicted age. The model can be formulated as Equation 1 with the training of  $g(X)$ .  $X$  is the input image features, the same set of features from the selected four models aforementioned.  $f(X)$  represents the PNC pre-trained model to compute  $\hat{y}$ , the estimated broad brain age.  $g(X)$  is a SVR module trained to explain the difference between  $\hat{y}$  and

chronological age, which presents the subtle brain age variation within a narrow age range.  $y$  is the final refined brain age.

$$\begin{aligned} \tilde{y} &= f(X); \dot{y} = g(X); y = \tilde{y} - \dot{y} \\ \text{Train: } \tilde{y} - \text{age} &\approx g(X) \end{aligned} \quad (1)$$

To train the refined models, we applied 10-fold cross-validation and grid-search hyper-parameter selection approach on 90% training data from ABCD baseline data. The performance of the refined brain age models was evaluated on 10% testing data of baseline, the full baseline data, and the full year-two follow-up data. Similarly, we only focused on the test performance on the full ABCD baseline, and year-two follow-up data to maintain the evaluation consistency among the brain age models.

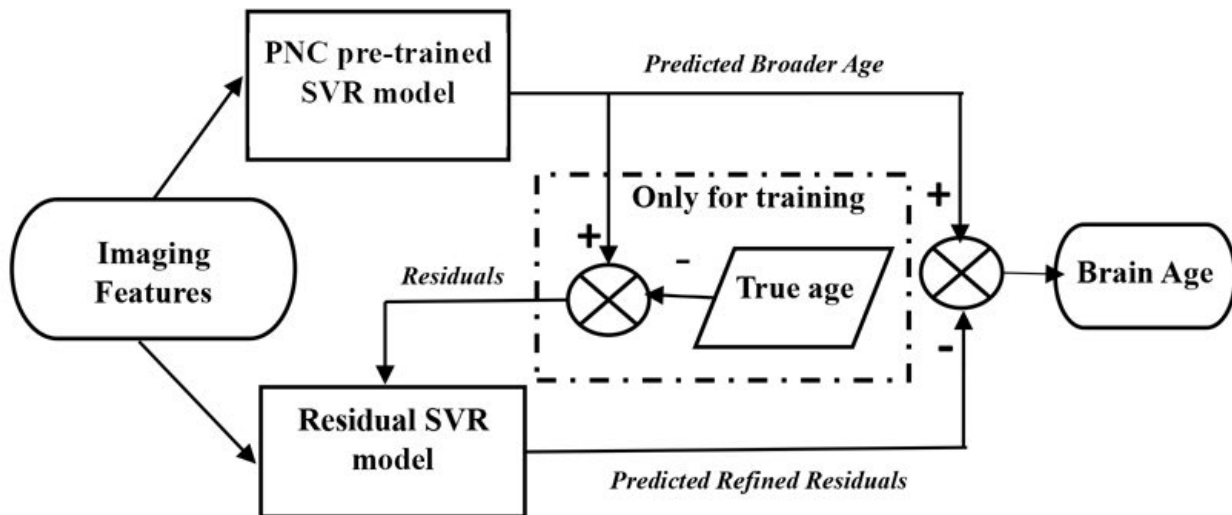


Figure 1: Refined brain age model.

### 3.4 Associations of brain age gap with cognitive measures

The brain age gap (the difference between the estimated brain age and the actual age) is calculated from the best-performing model. Then we examined the associations between individuals' brain age gap and their cognitive ability. Six NIH toolbox cognitive measures were investigated separately. They are scores from (i) Flanker attention task to test the speed and accuracy of participants' ability to identify congruent or incongruent targets given surrounding stimuli, (ii) pattern comparison processing speed task where the participants identify whether the pictures are the same or not from two given images and the score is calculated based on the number of correct answers, (iii) picture vocabulary task where the participants match the correct picture from the given four images by hearing an audio concept, (iv) picture sequence task where the participants replicate an activity sequence from a presented series of fifteen images that illustrate the particular activity, (v) the oral reading recognition task where the participants pronounce a single letter or word displayed in the middle of an iPad screen, and (vi) crystallized cognition composite score composed of picture vocabulary and oral reading recognition test.

These test scores indicate an individual's executive function or attention ability, visual information processing speed, verbal intellect, episodic memory or visuospatial sequencing memory, language/academic achievement, and composite crystallized intelligence. We used age-uncorrected standard scores to reflect age-related cognitive development. Linear mixed effect models were implemented with cognitive measures as the dependent variable, sex (male=0, female=1), age, and brain age gap as fixed effect independent variables, and ABCD family ID nested with ABCD site information as random effect variables. We included each subject's actual age and gender as predictors along with the brain age gap, as we want to control for the known age and gender effect on cognition during development.

For investigating the proportion of variance uniquely explained in the response variable (cognitive measure) by the fixed effect predictor (brain age gap), we have calculated adjusted  $R^2$  following the approach of Stoffel et al. [27]. The adjusted  $R^2$  is calculated by the difference between the variance

explained by the entire model, including all predictors, and the variance explained by a reduced model without the predictor of interest (brain age gap). We also reported a standardized coefficient value which is calculated by  $\beta \times (sd(x)/sd(y))$ , where  $\beta$  is the regression coefficient,  $x$  is the predictor and  $y$  is the response variable. Finally, we reported the p-values and corrected p-values (q-values) by multiple comparison corrections using the false discovery rate (FDR) method.

### 3.5 Important brain regions for age prediction

We have applied Shapley Additive Explanation (SHAP) technique [28] on our best-performing brain age model, refined *Combined\_RFE\_120* model, to interpret the features' contribution and to identify top brain areas which are crucial for brain age prediction. SHAP is a game theory-based method that computes the Shapley value for each feature by estimating the difference in the model outcome between a given sample and the averaged outcome across all data inputs. It reveals the individual feature contribution towards the model outcome for each sample or instance.

$$\sum_{j=1}^M \emptyset_j = f(x_i) - E_X(f(X)) \quad (2)$$

Here in the equation 2,  $x_i$  denotes one sample from a given dataset,  $X$  denotes the whole given dataset and  $\emptyset_j$  represents the SHAP value or feature importance of single feature  $j$ . The model outcome for a single sample is noted as  $f(x_i)$  and  $E_X(f(X))$  denotes averaged outcome value or expected value on the whole sample. The difference between the model outcome of one sample and the model expected value equals the sum of all feature importance for the sample.

For linear models, with independent features, we can approximate SHAP values directly from the model's weight coefficients. Given a linear model  $f(x) = \sum_{j=1}^M w_j x_j + b$ , the importance or SHAP value can be calculated as:  $\emptyset_j = w_j (x_j - E[X_j])$  where  $\emptyset_j$  is SHAP value,  $w_j$  and  $x_j$  represent regression coefficient and input value of feature  $j$  for the instance, respectively.  $E[X_j]$  refers to the expected value for  $X_j$ . We calculated the weights of the refined *Combined\_RFE\_120* model using the following formula.

$$\text{Broad age: } Y^{PNC} = \sum_{j=1}^M W_j^{PNC} \times X_j \quad (3)$$

$$\text{Residual age: } Y^{residual} = \sum_{j=1}^M W_j^{residual} \times X_j \quad (4)$$

$$\text{Refined brain age} = \sum_{j=1}^M (W_j^{PNC} - W_j^{residual}) \times X_j \quad (5)$$

Here in Equation 3,  $Y^{PNC}$ ,  $W^{PNC}$ , and  $X$  indicate the output, model weights, and 120 RFE features of the first module (PNC pre-trained model) of our refined model, respectively. In Equation 4,  $Y^{Residual}$  and  $W^{residual}$  correspond to the output and model weights of the second module (residual refinement) respectively. Finally, we calculated the refined model weights for each feature by subtracting the weights of the residual module from the weights of the PNC pre-trained module. We employed the linear SHAP method for the refined *Combined\_RFE\_120* model and approximated the SHAP value of each feature. The 120 features were ordered based on the mean absolute value of SHAP values across all subjects. We have identified the top eight contributing features for brain age prediction by considering the threshold value  $Z[mean(|SHAP|)] > 1.96$  (95% interval), where  $Z$  refers to the z-score of the mean absolute value of SHAP values. We then plotted the top eight important brain features for visual illustration purposes.

## 4 RESULTS

The performance of various brain age models is listed in Table 2 and Table 3 for ABCD baseline data and year-two follow-up data, respectively. The performance metrics include the mean of predicted brain age, correlation ( $r$ ) with predicted brain age and actual age, and MAE of predicted brain age.

Table 2 (top panel) from direct replication tests of PNC pre-trained models on ABCD baseline data showed that the FreeSurfer and combined pre-trained models gave relatively more accurate brain age prediction than the ICA and ROI models in terms of predicted mean age. ICA and ROI models resulted in either overestimated or underestimated brain age estimation. The predicted age from all models showed correlations with actual age ranging from 0.02 to 0.22, and MAE between 1.42 to 3.58 years. Considering participants of ABCD at baseline were all 9-10 years old, the relatively good-performed models, including *FreeSurfer\_152*, *FreeSurfer\_RFE\_110*, *Combined\_368*, and *Combined\_RFE\_120*, were selected for further testing on year-two follow-up. They were also selected as bases for further model building for independently newly trained ABCD models and for refined brain age models. Table 2 (middle panel) lists the results of ABCD self-trained models. The four models presented similar performance with a predicted mean age of 9.9 years old, correlation from 0.34 to 0.45, and MAE of 0.46 to 0.50. The *Combined\_368* model was slightly better than others. Table 2 (bottom panel) displayed the performance of four refined brain age models on the full baseline samples, where the model with combined features showed relatively better performance. The *Combined\_RFE\_120* model displayed the best performance with predicted age mean of 9.90, correlation of 0.36, and MAE of 0.49 years.

Table 2: Evaluation of PNC pre-trained brain age models (Top), ABCD self-trained brain age models (Middle), and refined brain age models (bottom) using ABCD Baseline data

Brain Age Models	Model Name	Results		
		Baseline (N= 11,382)		
		<i>Mean</i>	<i>Correlation (r)</i>	<i>MAE</i>
PNC Pre-trained	ICA_100	13.48	0.02	3.58
	ICA_RFE_81	12.75	0.03	2.89
	ROI_116	8.08	0.09	2.42
	ROI_RFE_70	7.70	0.08	2.66
	FreeSurfer_152	9.85	0.22	1.42
	FreeSurfer_RFE_110	9.76	0.22	1.43
	Combined_368	9.33	0.13	1.57
	Combined_RFE_120	9.53	0.13	1.56
ABCD self-trained	FreeSurfer_152	9.91	0.34	0.50
	FreeSurfer_RFE_110	9.90	0.33	0.50
	Combined_368	9.91	0.45	0.46
	Combined_RFE_120	9.90	0.38	0.49
Refined	FreeSurfer_152	9.89	0.17	0.76
	FreeSurfer_RFE_110	9.91	0.16	0.78
	Combined_368	9.91	0.21	0.60
	Combined_RFE_120	9.90	0.36	0.49

Table 3 presents the testing results on ABCD year-two follow-up data. Table 3 (top) showed the direct replication tests on four models with FreeSurfer and combined features, and all four models underestimated age at different levels, with mean age from 10.38 to 11.11. The correlation with actual age was from 0.16 to 0.27 and MAE from 1.56 to 1.92 years. Table 3 (middle) showed the performance of four ABCD self-trained models and all four models failed with various levels of underestimation of brain age. The predicted mean age was 10.38 to 11.11, and the MAE was 1.60 to 1.75. Table 3 (Bottom) listed the performance of refined models, and all four models achieved predicted mean age between 11.92 and 11.94 and MAE between 0.48 and 0.76. The best-performing model *Combined\_RFE\_120* produced a mean of 11.92, a correlation of 0.40, and an MAE of 0.48.

The best-performing model across baseline and follow-up data is the *Combined\_RFE\_120* refined

Table 3: Evaluation of PNC pre-Trained brain age models, ABCD self-trained brain age models, and Refined brain age Models using ABCD Year-Two Follow-Up data

Brain Age Models	Model Name	Results		
		Two years follow-up (N= 2877)		
		Mean	Correlation (r)	MAE
PNC Pre-trained	FreeSurfer_152	11.11	0.26	1.56
	FreeSurfer_RFE_110	10.88	0.27	1.64
	Combined_368	10.38	0.16	1.92
	Combined_RFE_120	10.88	0.16	1.69
ABCD self-trained	FreeSurfer_152	10.18	0.33	1.75
	FreeSurfer_RFE_110	10.18	0.33	1.75
	Combined_368	10.33	0.42	1.60
	Combined_RFE_120	10.22	0.37	1.72
Refined	FreeSurfer_152	11.94	0.20	0.76
	FreeSurfer_RFE_110	11.93	0.21	0.74
	Combined_368	11.93	0.26	0.58
	Combined_RFE_120	11.92	0.40	0.48

Table 4: Associations of brain age gap from the refined *Combined\_RFE\_120* model with cognitive measures at baseline and year-two follow-up.

Neurocognitive Assessments (Age uncorrected score)	Baseline (N= 11,382)				Two years follow-up (N= 2877)			
	Standardized Coefficient	Effect size (Adjusted R <sup>2</sup> )	P	Adjusted P (q-value)	Standardized Coefficient	Effect size (Adjusted R <sup>2</sup> )	P	Adjusted p (q-value)
Flanker Task	1.90E-2	2.0E-4	2.92E-1	2.92E-1	2.42E-2	3.0E-4	4.50E-1	5.26E-1
Pattern Comparison Processing Speed	9.96E-2	3.1E-3	2.13E-08	<b>1.28E-07</b>	7.29E-2	2.2E-3	2.13E-2	1.279E-1
Picture Vocabulary Task	6.40E-2	1.4E-3	1.87E-4	<b>5.61E-04</b>	5.9E-2	1.5E-3	5.08E-2	1.524E-1
Picture Sequence Memory	3.75E-2	5.0E-4	3.9E-2	5.17E-2	2.71E-2	3.0E-4	3.9907E-1	5.26E-1
Oral Reading Recognition Task	3.53E-2	5.0E-4	4.31E-2	5.17E-2	2.0E-2	2.0E-4	5.26E-1	5.26E-1
Crystallized Intelligence Composite	5.83E-2	1.2E-3	5.54E-4	<b>1.1E-3</b>	5.03E-2	1.2E-3	9.39E-2	1.878E-1

note: significant values ( $q \leq 0.05$ ) are shown in bold.

model. Therefore, focusing on the *Combined\_RFE\_120* refined model, we tested the relationship of the brain age gap with cognitive measures at baseline and year-two follow-up. Table 4 lists the standardized coefficient value, adjusted  $R^2$ , p-values, and adjusted p-values (q-values) from association tests via linear mixed models. The significant results passing FDR correction are listed in bold. The positive coefficient value indicates that the subject's cognitive ability increases with accelerated brain age.

We have also identified the top 8 contributing features of *Combined\_RFE\_120* refined model using SHAP values as plotted in Figure 2. Figure 2 (a) shows a beeswarm plot indicating the top eight contributing features' impact on brain age prediction. Each dot denotes the SHAP value of one feature for one individual subject, where the SHAP value is along the x-axis. and the blue-purple color indicates the input feature value from low to high. Figure 2 (b,c) shows the brain areas involved in the selected top contributing features in the coronal and sagittal view, color-coded as listed in Figure 2 (d).

## 5 Discussion and Conclusion

Recently many studies have used regression-based machine learning approaches or deep learning models to estimate brain age with high precision [29–32]. Very few studies focus on analyzing the generalizability of brain age prediction models. The study [32] shows that unseen data and different age ranges of samples can influence the generalizability of the brain age model. In this study, we evaluated

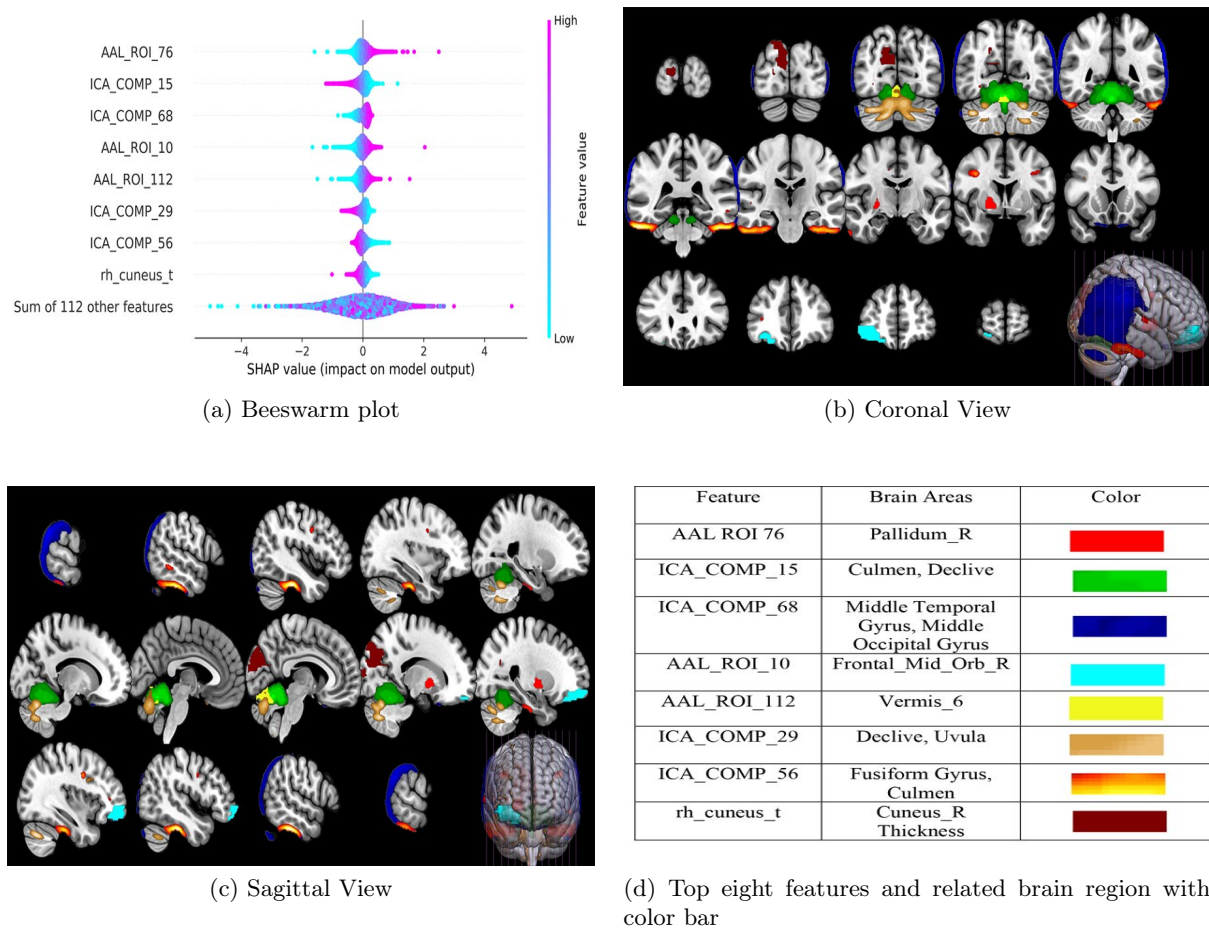


Figure 2: Top eight features and the related brain areas from the refined model. a) The beeswarm plots showing the contribution (SHAP value) of the top eight features. b) Coronal view of top eight brain regions. c) Sagittal view of the top eight brain regions. d) Color code for the top eight features and their related brain locations.

the generalizability of the pre-trained brain age models (trained on PNC subjects with a broader age range of 8-22 years) from our previous study [1] by applying them to independent ABCD baseline and year-two follow-up data of children aged between 9-10 or 11-12 years. Our analysis showed that the model with only ICA component features or mean gray matter density features of brain regions showed lower accuracies in age prediction relative to the Freesurfer features. The models with ICA features produced overestimated predictions with low correlations with the actual age and high MAEs. In contrast, the models with mean gray matter density of ROIs underestimated the brain age with similar low correlations and high MAEs. We have followed the same preprocessing pipeline for both PNC and ABCD structural MRI images. However, the different group mean age of the subjects (approx. 16 years for PNC data, 9.9 and 11.93 years for ABCD baseline and year-two data, respectively) of PNC and ABCD cohorts might have induced some variability during the image spatial-normalization phase of preprocessing steps. That might be a possible reason behind the low performance of the ICA and AAL gray matter density ROI models. We have observed that the models with FreeSurfer and combined features achieved relatively better performance, with predicted brain age being closer to the mean of actual age from 9.33 to 9.85 at baseline and 10.38 to 11.11 at year-two follow-up. But the MAE of predicted age was still high, from 1.42 to 1.57 at baseline and 1.56 to 1.92 at year-two, and the predicted age did not correlate strongly with the actual age on both baseline and year-two follow-up.

Focusing on Freesurfer and combined features, we trained the new ABCD brain age models using the ABCD baseline data to test the prediction power of the ABCD features. As expected, the ABCD self-trained brain age models showed improved accuracy when tested on baseline data, but the performance decreased when tested on the year-two follow-up data. Basically, all four ABCD self-trained brain age models failed while tested on year-two follow-up data when participants were 11-12 years old. For

instance, the performance of the *Combined*\_368 model declined markedly with the MAE increasing from 0.46 at baseline to 1.92 at year-two follow-up.

To improve the generalizability of the brain age model, we have constructed refined brain age prediction approach. Our results indicate that the refined brain age models showed better prediction accuracy compared to both pre-trained models and self-trained models. In particular, for the *Combined*\_RFE\_120 refined model, the correlation with actual age significantly improved from 0.13 to 0.36, and the MAE decreased from 1.56 to 0.49 years for the baseline data (Table II). Similarly, we also observed big improvements for the year-two follow-up data for the *Combined*\_RFE\_120 refined model, when the correlation with actual age increased from 0.16 to 0.40 and the MAE decreased from 1.69 to 0.48 years (see Table 3). Unlike ABCD self-trained models, refined brain age models achieved improved performance on both baseline and year-two data.

Overall, the PNC pre-trained models failed to produce valid age prediction for both ABCD baseline and year-two follow-up, the ABCD self-trained models performed very well at baseline but failed at year-two follow-up, while the ABCD refined brain age models showed consistently good performance at both baseline and year-two follow-up. Even though all three types of models are linear, the weights are trained with different loss functions: minimize errors in a large age range, minimize errors in a narrow age range, and minimize errors in an adjusted narrow age range. The adjusted ages prevented the model to be over-fitted to the narrow age range.

For interpreting the meaning of the brain age gap, we have applied linear mixed-effect models to test the associations between the derived brain age gap and six neurocognitive measures. As expected, age had a significant impact on all six cognitive measures for both time points, and sex had a significant influence on episodic memory and information processing speed at baseline and year-two follow-up. Detailed results are not presented here since age and sex effect on cognition is outside the scope of this study. Beyond age and sex, there is a significant association between the brain age gap and the rate of information processing speed (pattern comparison), verbal comprehension (picture vocabulary), and crystallized intelligence on baseline data. The adjusted  $R^2$  score, partial variance explained is 0.0031, 0.0014, and 0.0012, respectively, indicating accelerated brain age explains partially, small yet significant, the advance in cognitive ability. For the year-two analysis, the uncorrected p-values were less than 0.05 for pattern comparison and 0.05 for the picture vocabulary task. Even though the association tests produced similar effect sizes with consistent positive effects between baseline and year-two follow-up, we did not observe any significant influence (based on FDR corrected q-value) of the brain age gap over the cognitive measures for the year-two data. It may be due to a comparatively smaller sample size of data (baseline: N=11,382, year-two: N=2877) in the year-two study.

Eight brain features are selected to be important for brain age prediction. Based on the features' SHAP values (Figure 2(a)), we can interpret how the top features influence brain age. For example, the highest contributing feature, *AAL\_ROI\_76*, has SHAP values ranging from about -3 to 3. The positive SHAP values are presented in purple color, indicating higher input feature values, while negative SHAP values are in blue indicating lower input features. Thus, a higher value of the feature *AAL\_ROI\_76* relates to an increased brain age. Vice versa, a lower feature value in *AAL\_ROI\_76* links to a decreased brain age. *AAL\_ROI\_76* is the gray matter volume of the right Pallidum shown in Figure 2(b,c). Therefore, our model suggests an increased gray matter density in the right pallidum area is a biomarker for the subject's brain development/maturation during this sensitive developing phase. Similarly, the relationship between the SHAP value and the input value of the second top feature (*ICA\_COMP\_15* positive SHAP values link to low input values) suggests lower gray matter density in the culmen area is associated with increased brain age. In sum, our results suggest that the higher gray matter density in the right pallidum, right middle orbital frontal gyrus, middle temporal gyrus, and vermis is associated with increased brain age, while the lower gray matter density in the culmen, declive and fusiform gyrus, as well as reduced cortical thickness of Cuneus, is associated with increased brain age. It is well documented that gray matter develops in a region-specific manner with a 'posterior to anterior' maturation order at large [33], and presents both linear and non-linear (e.g., inverted U shape) developmental trajectories [34–36]. For instance, Wierenga and colleagues [37] focused on subcortical regions' developmental trajectory and reported that pallidum presents a curvilinear trajectory peaked at 17 years old. Thus, at the age of 9 to 12, the increase

of pallidum volume in the participants of the ABCD cohort indicates continued brain development and increased brain age. For the cortical area of the brain abundant evidence supports that higher-order association cortices mature only after lower-order somatosensory and visual cortices, with gray matter thickness peaking at different ages [33, 38]. At the age of 9-12, our model indicates brain development is linked to reduced gray matter density and thickness in the occipital lobe (fusiform gyrus, cuneus) and increased gray matter density in the frontal and temporal lobe, suggesting the visual cortex has passed its maturation peak but the higher order frontal and temporal cortices are still approaching to their peaks. In contrast, few studies have investigated cerebellum and cerebellar subregions development during childhood or life span [39, 40]. Consistently both studies identified that most of the cerebellar subregions followed an inverted U-shaped developmental trajectory and Romero et al. also highlighted the Lobule IX (the uvula) showed the fastest development, reaching the maximum at around 5 years [40]. Our model also makes a note of gray matter density in Uvula: reduced gray matter in Uvula indicates increased brain age at the age of 9-12. In addition, our model also grouped culmen, declive, and uvula, to have similar development change at this period of age. Overall, our findings indicate that in the pre-teens age, while cerebellum and occipital regions start gray matter reduction, the frontal and temporal regions are still in the increasing phase, which is in line with known brain development trajectory in children and adolescents [41, 42].

In summary, our study shows that the PNC pre-trained models have prediction power to capture the overall mean age of ABCD samples but cannot precisely estimate the brain age within the narrow range. In contrast, ABCD self-trained brain age models can only give a precise estimation of the subjects between 9-10 years but cannot be generalized to other age ranges. Our proposed refined age methods utilize the broader prediction power of pre-trained models and refined the model with residual to increase the overall robustness of the brain age models. The strength of this study should be interpreted in light of the following limitations. The two datasets tested (ABCD baseline and year-two follow-up) have an age range of pre-teens, and the robustness of the refined model requires more tests using data from teenagers and young adulthood, as well as data from independent samples other than ABCD. As the brain develops in a nonlinear manner during childhood and adolescence, the effectiveness of refined models may be subjected to periods of development, such as childhood, pre-teens, teenagers, and young adulthood. Future evaluation is warranted.

## Acknowledgment

This work is supported by NIH R01DA049238 and NSF 2112455 to VC and JL.

## References

- [1] B. Ray, K. Duan, J. Chen, Z. Fu, P. Suresh, S. Johnson, V. D. Calhoun, and J. Liu, "Multimodal brain age prediction with feature selection and comparison," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3858–3864, IEEE, 2021.
- [2] S. Basodi, R. Raja, B. Ray, H. Gazula, J. Liu, E. Verner, and V. D. Calhoun, "Federation of brain age estimation in structural neuroimaging data," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3854–3857, IEEE, 2021.
- [3] S. Basodi, R. Raja, B. Ray, H. Gazula, A. D. Sarwate, S. Plis, J. Liu, E. Verner, and V. D. Calhoun, "Decentralized brain age estimation using mri data," *Neuroinformatics*, pp. 1–10, 2022.
- [4] N. Egorova, F. Liem, V. Hachinski, and A. Brodtmann, "Predicted brain age after stroke," *Frontiers in aging neuroscience*, vol. 11, p. 348, 2019.
- [5] I. Beheshti, S. Mishra, D. Sone, P. Khanna, and H. Matsuda, "T1-weighted mri-driven brain age estimation in alzheimer's disease and parkinson's disease," *Aging and disease*, vol. 11, no. 3, p. 618, 2020.

- [6] H. Liang, F. Zhang, and X. Niu, “Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders,” tech. rep., Wiley Online Library, 2019.
- [7] J. H. Cole and K. Franke, “Predicting age using neuroimaging: innovative brain ageing biomarkers,” *Trends in neurosciences*, vol. 40, no. 12, pp. 681–690, 2017.
- [8] J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector, and G. Montana, “Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker,” *NeuroImage*, vol. 163, pp. 115–124, 2017.
- [9] J. H. Cole, S. J. Ritchie, M. E. Bastin, V. Hernández, S. Muñoz Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, *et al.*, “Brain age predicts mortality,” *Molecular psychiatry*, vol. 23, no. 5, pp. 1385–1392, 2018.
- [10] K. Franke, E. Luders, A. May, M. Wilke, and C. Gaser, “Brain maturation: predicting individual brainage in children and adolescents using structural mri,” *Neuroimage*, vol. 63, no. 3, pp. 1305–1312, 2012.
- [11] G. Erus, H. Battapady, T. D. Satterthwaite, H. Hakonarson, R. E. Gur, C. Davatzikos, and R. C. Gur, “Imaging patterns of brain development and their relationship to cognition,” *Cerebral cortex*, vol. 25, no. 6, pp. 1676–1684, 2015.
- [12] F. Liem, G. Varoquaux, J. Kynast, F. Beyer, S. K. Masouleh, J. M. Huntenburg, L. Lampe, M. Rahim, A. Abraham, R. C. Craddock, *et al.*, “Predicting brain-age from multimodal imaging data captures cognitive impairment,” *Neuroimage*, vol. 148, pp. 179–188, 2017.
- [13] M. J. Lund, D. Alnæs, A.-M. G. de Lange, O. A. Andreassen, L. T. Westlye, and T. Kaufmann, “Brain age prediction using fmri network coupling in youths and associations with psychiatric symptoms,” *NeuroImage: Clinical*, vol. 33, p. 102921, 2022.
- [14] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, “Accurate brain age prediction with lightweight deep neural networks,” *Medical image analysis*, vol. 68, p. 101871, 2021.
- [15] P. Sturmfels, S. Rutherford, M. Angstadt, M. Peterson, C. Sripada, and J. Wiens, “A domain guided cnn architecture for predicting age from structural brain images,” in *Machine learning for healthcare conference*, pp. 295–311, PMLR, 2018.
- [16] J. Hong, Z. Feng, S.-H. Wang, A. Peet, Y.-D. Zhang, Y. Sun, and M. Yang, “Brain age prediction of children using routine brain mr images via deep learning,” *Frontiers in Neurology*, vol. 11, p. 584682, 2020.
- [17] X. Feng, Z. C. Lipton, J. Yang, S. A. Small, F. A. Provenzano, A. D. N. Initiative, F. L. D. N. Initiative, *et al.*, “Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging,” *Neurobiology of aging*, vol. 91, pp. 15–25, 2020.
- [18] B. A. Jónsson, G. Bjornsdottir, T. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. Gudbjartsson, H. Stefansson, K. Stefansson, and M. Ulfarsson, “Brain age prediction using deep learning uncovers associated sequence variants,” *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.
- [19] M. Ueda, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki, “An age estimation method using 3d-cnn from brain mri images,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 380–383, IEEE, 2019.
- [20] T. D. Satterthwaite, J. J. Connolly, K. Ruparel, M. E. Calkins, C. Jackson, M. A. Elliott, D. R. Roalf, R. Hopson, K. Prabhakaran, M. Behr, *et al.*, “The philadelphia neurodevelopmental cohort: A publicly available resource for the study of normal and abnormal brain development in youth,” *Neuroimage*, vol. 124, pp. 1115–1119, 2016.

- [21] D. M. Barch, M. D. Albaugh, S. Avenevoli, L. Chang, D. B. Clark, M. D. Glantz, J. J. Hudziak, T. L. Jernigan, S. F. Tapert, D. Yurgelun-Todd, *et al.*, “Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: Rationale and description,” *Developmental cognitive neuroscience*, vol. 32, pp. 55–66, 2018.
- [22] N. D. Volkow, G. F. Koob, R. T. Croyle, D. W. Bianchi, J. A. Gordon, W. J. Koroshetz, E. J. Pérez-Stable, W. T. Riley, M. H. Bloch, K. Conway, *et al.*, “The conception of the abcd study: From substance use to a broad nih collaboration,” *Developmental cognitive neuroscience*, vol. 32, pp. 4–7, 2018.
- [23] D. J. Hagler Jr, S. Hatton, M. D. Cornejo, C. Makowski, D. A. Fair, A. S. Dick, M. T. Sutherland, B. Casey, D. M. Barch, M. P. Harms, *et al.*, “Image processing and analysis methods for the adolescent brain cognitive development study,” *Neuroimage*, vol. 202, p. 116091, 2019.
- [24] L. Xu, K. M. Groth, G. Pearlson, D. J. Schretlen, and V. D. Calhoun, “Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia,” *Human brain mapping*, vol. 30, no. 3, pp. 711–724, 2009.
- [25] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, “Automated anatomical labelling atlas 3,” *Neuroimage*, vol. 206, p. 116189, 2020.
- [26] A. R. Khan, L. Wang, and M. F. Beg, “Freesurfer-initiated fully-automated subcortical brain segmentation in mri using large deformation diffeomorphic metric mapping,” *Neuroimage*, vol. 41, no. 3, pp. 735–746, 2008.
- [27] M. A. Stoffel, S. Nakagawa, and H. Schielzeth, “part2: partitioning r<sup>2</sup> in generalized linear mixed models,” *PeerJ*, vol. 9, p. e11414, 2021.
- [28] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [29] S. Valizadeh, J. Hänggi, S. Mérillat, and L. Jäncke, “Age prediction on the basis of brain anatomical measures,” *Human brain mapping*, vol. 38, no. 2, pp. 997–1008, 2017.
- [30] X. Niu, F. Zhang, J. Kounios, and H. Liang, “Improved prediction of brain age using multimodal neuroimaging data,” *Human brain mapping*, vol. 41, no. 6, pp. 1626–1643, 2020.
- [31] H. Jiang, N. Lu, K. Chen, L. Yao, K. Li, J. Zhang, and X. Guo, “Predicting brain age of healthy adults based on structural mri parcellation using convolutional neural networks,” *Frontiers in neurology*, vol. 10, p. 1346, 2020.
- [32] A.-M. G. de Lange, M. Anatürk, J. Rokicki, L. K. Han, K. Franke, D. Alnæs, K. P. Ebmeier, B. Draganski, T. Kaufmann, L. T. Westlye, *et al.*, “Mind the gap: Performance metric evaluation in brain-age prediction,” *Human Brain Mapping*, 2022.
- [33] N. Gogtay, J. N. Giedd, L. Lusk, K. M. Hayashi, D. Greenstein, A. C. Vaituzis, T. F. Nugent III, D. H. Herman, L. S. Clasen, A. W. Toga, *et al.*, “Dynamic mapping of human cortical development during childhood through early adulthood,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 21, pp. 8174–8179, 2004.
- [34] P. Shaw, N. J. Kabani, J. P. Lerch, K. Eckstrand, R. Lenroot, N. Gogtay, D. Greenstein, L. Clasen, A. Evans, J. L. Rapoport, *et al.*, “Neurodevelopmental trajectories of the human cerebral cortex,” *Journal of neuroscience*, vol. 28, no. 14, pp. 3586–3594, 2008.
- [35] J. N. Giedd, J. Blumenthal, N. O. Jeffries, F. X. Castellanos, H. Liu, A. Zijdenbos, T. Paus, A. C. Evans, and J. L. Rapoport, “Brain development during childhood and adolescence: a longitudinal mri study,” *Nature neuroscience*, vol. 2, no. 10, pp. 861–863, 1999.

- [36] T. Paus, “Mapping brain maturation and cognitive development during adolescence,” *Trends in cognitive sciences*, vol. 9, no. 2, pp. 60–68, 2005.
- [37] L. Wierenga, M. Langen, S. Ambrosino, S. van Dijk, B. Oranje, and S. Durston, “Typical development of basal ganglia, hippocampus, amygdala and cerebellum from age 7 to 24,” *Neuroimage*, vol. 96, pp. 67–72, 2014.
- [38] L. M. Wierenga, M. Langen, B. Oranje, and S. Durston, “Unique developmental trajectories of cortical thickness and surface area,” *Neuroimage*, vol. 87, pp. 120–126, 2014.
- [39] H. Tiemeier, R. K. Lenroot, D. K. Greenstein, L. Tran, R. Pierson, and J. N. Giedd, “Cerebellum development during childhood and adolescence: a longitudinal morphometric mri study,” *Neuroimage*, vol. 49, no. 1, pp. 63–70, 2010.
- [40] J. E. Romero, P. Coupe, E. Lanuza, G. Catheline, J. V. Manjón, and A. D. N. Initiative, “Toward a unified analysis of cerebellum maturation and aging across the entire lifespan: a mri analysis,” *Human Brain Mapping*, vol. 42, no. 5, pp. 1287–1303, 2021.
- [41] S. Frangou, A. Modabbernia, S. C. Williams, E. Papachristou, G. E. Doucet, I. Agartz, M. Aghajani, T. N. Akudjedu, A. Albajes-Eizagirre, D. Alnæs, *et al.*, “Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3–90 years,” *Human brain mapping*, vol. 43, no. 1, pp. 431–451, 2022.
- [42] G. Ball, J. Seidlitz, R. Beare, and M. Seal, “Cortical remodelling in childhood is associated with genes enriched for neurodevelopmental disorders. *neuroimage*, 215, article 116803,” 2020.