Systematic Evaluation of Machine Learning Algorithms for Neuroanatomically-Based Age Prediction in Youth

Amirhossein Modabbernia[1], Heather C. Whalley[2], David C. Glahn[3], Paul M. Thompson[4], Rene S. Kahn[1], Sophia Frangou[1.5]

1 Department of Psychiatry, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA

2 Division of Psychiatry, University of Edinburgh, Kennedy Tower, Royal Edinburgh Hospital, Morningside Park, Edinburgh, EH10 5HF, UK

3 Boston Children's Hospital and Harvard Medical School, 300 Longwood Avenue Boston, MA 02115, USA

4 Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Marina del Rey, 2025 Zonal Avenue, Los Angeles, CA 90033, USA

5. Department of Psychiatry, Djavad Mowafaghian Centre for Brain Health, University of British Columbia, 2215 Wesbrook Mall, Vancouver, BC V6T 1Z3, Canada

Corresponding author: Sophia Frangou, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, 1 Gustave L Levy Place, New York NY 10029

Email: sophia.frangou@mssm.edu

Highlights

- **Ensemble-based algorithms performed best in predicting brain age during development**
- **Support vector regression offers optimal prediction accuracy and computational costs**
- **A 400-parcel resolution provided the best accuracy and computational efficiency**

Abstract

Application of machine learning algorithms to structural magnetic resonance imaging (sMRI) data has yielded behaviorally meaningful estimates of the biological age of the brain (brain-age). The choice of the machine learning approach in estimating brain-age in children and adolescents is important because age-related brain changes in these age-groups are dynamic. However, the comparative performance of the multiple machine learning algorithms available has not been systematically appraised. To address this gap, the present study evaluated the accuracy (Mean Absolute Error; MAE) and computational efficiency of 21 machine learning algorithms using sMRI data from 2,105 typically developing individuals aged 5 to 22 years from five cohorts. The trained models were then tested in an independent holdout datasets, comprising 4,078 pre-adolescents (aged 9-10 years). The algorithms encompassed parametric and nonparametric, Bayesian, linear and nonlinear, tree-based, and kernel-based models. Sensitivity analyses were performed for parcellation scheme, number of neuroimaging input features, number of cross-validation folds, and sample size. The best performing algorithms were Extreme Gradient Boosting (MAE of 1.25 years for females and 1.57 years for males), Random Forest Regression (MAE of 1.23 years for females and 1.65 years for males) and Support Vector Regression with Radial Basis Function Kernel (MAE of 1.47 years for females and 1.72 years for males) which had acceptable and comparable computational efficiency. Findings of the present study could be used as a guide for optimizing methodology when quantifying age-related changes during development.

Keywords: machine learning; development; brain-based age prediction; youth

1. Introduction

Brain development involves highly organized multistep processes (Tau and Peterson, 2010) that lead to the emergence of adult levels of cognitive and behavioral competency (Paus, 2005; Spear, 2000). Brain development involves numerous cellular and noncellular events (Tau and Peterson, 2010), which are below the resolution of magnetic resonance imaging (MRI) but underpin morphological changes in brain organization that can be captured using structural MRI (sMRI) techniques. Multiple studies have shown that the volume of subcortical structures typically peaks in late childhood and adolescence and decreases thereafter (Dima et al., 2021; Raznahan et al., 2014). Cortical thickness shows a steep reduction in late childhood and adolescence that continues at a slower rate throughout adult life (Frangou et al., 2021; Wierenga et al., 2020). Cortical surface area expands during childhood and most of adolescence showing gradual decrements thereafter (Fjell et al., 2015; Tamnes et al., 2017).

These age-related changes demonstrate marked inter-regional and inter-individual variation (Mills et al., 2021; Wierenga et al., 2020).

Machine learning (ML) algorithms applied to sMRI data can harness the multidimensional nature of age-related brain changes at the level of each individual subject to predict age, as a proxy for the biological age of the brain (i.e., brain-age). The difference between brain-age and chronological age is referred to here as brain-age-gap-estimation (BrainAGE)(Franke and Gaser, 2019), which is equivalent to terms such as brain-predicted-age-difference (brainPAD) (Luna et al., 2021), brain-age-gap (BAG) (Anaturk et al., 2021) and brain-age delta(Beheshti et al., 2019) used in other studies. In adults, higher brain-age relative to chronological age (i.e., higher BrainAGE) has been associated with adverse physical (Cole et al., 2018), cognitive (Anaturk et al., 2021; Boyle et al., 2021; Elliott et al., 2019) and mental health phenotypes (Kaufmann et al., 2019; Lee et al., 2021). By contrast, in children and adolescents higher BrainAGE has been associated with better cognitive test

performance (Boyle et al., 2021; Erus et al., 2015; Luna et al., 2021) while associations with clinical phenotypes show a more complex pattern which may depend on the nature of the phenotype and/or the developmental stage of the sample (Chung et al., 2018; Luna et al., 2021). These findings underscore the importance of accuracy in brain-based age prediction in youth, as childhood and adolescence are arguably periods of the most dynamic brain re-organization.

The functional significance of brainAGE in youth is outside the scope of this study. Here we focus exclusively on the evaluation of the methods used to compute brain-age from sMRI data.

With few exceptions (Ball et al., 2021; Brouwer et al., 2021; Lee et al., 2021; Luna et al., 2021), studies on brain-age prediction have typically employed a single ML algorithm, most commonly relevance vector regression (RVR), Gaussian process regression (GPR), and support vector regression (SVR) (Cole et al., 2018; Franke et al., 2012; Franke et al., 2010; Gaser et al., 2013; Liem et al., 2017; Valizadeh et al., 2017). We have previously shown in adults that brain-based age prediction from sMRI data is substantially influenced by the choice of algorithm (Lee et al., 2021). Here we address the challenge of benchmarking brain-based age prediction methods in youth by systematically evaluating 21 ML algorithms on sMRI data from five different cohorts and then testing their performance in an independent sample. The algorithms encompassed parametric and nonparametric, Bayesian, linear and nonlinear, tree-based, and kernel-based models. These algorithms were selected to include those that are commonly used in brain-age prediction studies as well representative examples of a range of algorithms that provide reasonable and potentially better alternatives. Beyond model accuracy, our evaluations extended to testing the ML methods for their sensitivity to key parameters known to affect accuracy which pertain to parcellation scheme, number of neuroimaging input features (Valizadeh et al., 2017), number of cross-validation folds and sample size (by resampling the available data). Our initial prediction was that nonlinear kernel-based and ensemble algorithms would

outperform other algorithms because they are theoretically better at handling colinear data and non-linear relationships with age and, in the case of ensemble algorithms, they improve predictive performance by aggregating results from multiple nodes. Collectively, these analyses may assist in optimizing the design of future investigations on brain predicted age in youth.

2. Method

2.1 Samples

We used T1-weighted scans from six separate cohorts: the Autism Brain Imaging Data Exchange (ABIDE) (Di Martino et al., 2017; Di Martino et al., 2014); ABIDE II (Di Martino et al., 2017); ADHD-200 (Consortium, 2012); Human Connectome Project Development (HCP-D) (Harms et al., 2018); Child Mind Institute (CMI) (Alexander et al., 2017), Adolescent Brain Cognitive Development (ABCD) (Alexander et al., 2017; Garavan et al., 2018) (supplemental material). Data collection for these cohorts was conducted at multiple independent sites located in eight countries: (USA, Germany, Ireland, Belgium, the Netherlands, Switzerland, China, and France). Only psychiatrically healthy participants with high-quality anatomical brain scans from each cohort were included (Supplemental Material and Supplemental Table 1). Data from five cohorts (total n=2,105; 41% female, Supplemental Figure 1) were used to train the ML algorithms (training set) while data from the ABCD sample (n=4,078; 52% female) comprised the independent hold-out test-sets.

2.2 Image processing

Across all cohorts, more than 98% of the participants were scanned using 3-T MRI machines; Siemens Prisma and Trio Tim scanners were each used for 31% of the participants of the total training sample (Supplemental Table 1)**.** The T1-weighted images were downloaded from the respective cohort repositories and processed at the Icahn School of Medicine at Mount Sinai (ISMMS) using identical pipelines. Image processing was implemented using standard pipelines in the FreeSurfer 7.1.0 software to generate cortical parcels based on the Schaefer (Schaefer et al., 2018) scheme by projecting the parcellation onto individual surface space (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/project_to_individual) and using the *mri_anatomical_stats* function to extract cortical values. We used the 400-parcel resolution (i.e., 400 cortical thickness and 400 cortical surface area values) in the main analyses

(Supplemental Figure 2). The 400-parcellation scheme has been widely studied with respect to its stability, signal to noise ratio, performance in different contexts and correspondence to histology (Bryce et al., 2021; Valk et al., 2020). Participants with missing values on any parcellations were excluded, because it was assumed that in the case of missing value the image quality was compromised; participants were also excluded if more than 5% of their parcellation features had extreme values (details in supplemental material).

2.3 Algorithms for brain-based age prediction

We used the *caret* package (version 6.0.84) in R (version 3.5.3) to conduct the ML analyses because it interfaces with multiple ML packages and standardizes data preprocessing, model training and testing. Several regression algorithms can be extended to accommodate non-linear associations using kernel functions. A kernel function generally transforms the original non-linear data into a higher-dimensional space in which they can become linearly separable. The kernelized models evaluated here incorporated polynomial and radial basis function (RBF) kernels. The former adds features using the polynomial combinations of the original data up to a specified degree and the latter adds features using the distance of the original data from specified reference values. Below we describe the 15 base models, six of which have non-linear kernelized variations; together, they amount to 21 different algorithms:

(1)**Generalized linear model:** This is a standard algorithm for regression that minimizes the sum of squared errors between the observed variables and predicted outcomes. Models have no tuning parameters and were implemented using the **"**glm" function.

(2)**Bayesian general linear model** (Gelman et al., 2008)**:** This a linear regression model in which the outcome variable and the model parameters are assumed to be drawn from a probability distribution; it therefore provides estimates of model uncertainty. Models have no tuning parameters and were implemented using the "bayesglm" function.

(3)**Gaussian Processes Regression** (Williams and Barber, 1998): This is a regression model that follows Bayesian principles. The covariance function here was defined by using either a linear function, or a polynomial or an RBF kernel as a prior. The

polynomial kernels were tuned using degree and scale and the RBF kernels were tuned using the *sigma* parameter (the inverse kernel width parameter). Models were implemented using “gaussprRadial”, “gaussprLinear”, and “gaussprPoly”.

(4)**Independent Component Regression** (Shao et al., 2006)**:** This is a linear regression model in which components from a prior independent component analysis are used as the explanatory variables. The number of components was tuned, and the models were implemented using the “icr” function.

(5)**Principal Component Regression**: This is a linear regression model in which components from a prior principal component analysis are used as the explanatory variables. The number of components was tuned, and the models were implemented using the “pcr” function.

(6)**Kernel Partial Least Squares Regression** (Dayal and MacGregor, 1997): This is an extension of the partial least squares (PLS) regression which creates components by using the correlations between explanatory variables and outcome variables. The kernelized version used here (K-PLS) maps the data vector from the sample space to a higher-dimensional, Euclidean space; models were tuned for the number of components and implemented using the ”kernelpls” function.

(7)**Sparse Partial Least Squares Regression (SPLS)**(Chun and Keleş, 2010): This is a different extension of PLS that reduces the number of explanatory variables (sparsity) through a least absolute shrinkage and selection operator (LASSO) approach. The models were tuned for the number of components, and *eta* (the sparsity parameter), and were implemented using the “spls” function.

(8)**Quantile Regression with least absolute shrinkage and selection operator (LASSO) Penalty** (Wu and Liu, 2009): This algorithm models the relationship between explanatory variables and specific percentiles (or "quantiles") of the outcome variable; in this variation, sparsity was introduced through the LASSO approach. The number of selected variables was tuned, and models were implemented with the “rqlasso” function.

(9)**Elastic Net Regression** (Zou and Hastie, 2005)**:** This is a linear regression that adds two penalties, LASSO regression (L1-norm) and ridge regression (L2-norm), in the loss function to encourage simpler models and avoid overfitting. Models were tuned for *lambda* (weight decay) and fraction of the full solution (equivalent to ordinary least squares) and were implemented using the "enet" function.

(10)**Boosted Generalized Additive Model** (Bühlmann and Yu, 2003)**:** This generalized additive model is fitted using a gradient-based boosting algorithm based on penalized B-splines. Overfitting was reduced by pruning the number of iterations using the optimal value of the Akaike Information Criterion. Models were implemented using the "gamboost" function.

(11)**Random Forest Regression** (Breiman, 2001)**:** This an ensemble machine learning method, which involves construction of multiple decision trees (i.e., forests) via bootstrap (bagging) and aggregates the predictions from these multiple trees to reduce the variance and improve the robustness and precision of the results. Models were implemented using the "rf" function and were tuned with regard to the number of trees.

(12)**Support Vector Regression** (Cortes and Vapnik, 1995): Support Vector Regression (SVR) is characterized by the use of kernels, sparsity, and control of the margin of tolerance (epsilon; ε) and the number of support vectors (Awad and Khanna, 2015). It identifies a symmetrical ε-insensitive region, called the ε-tube, which approaches the loss function as an optimization problem; the ε-value determines the width of the tube and maximization of the "flatness" aims to ensure that it contains most of the values in the training sample. Here flatness maximization was subject to the L2-norm penalty. In addition to the linear kernel, we also tested a version with polynomial and RBF kernels. The corresponding functions were "svmLinear3", "svmPoly" and "svmRadial". The regularization parameter (C) was used to optimize all models, while scale and degree were also considered in polynomial models and *sigma* for RBF models.

(13)**Relevance Vector Regression** (Tipping, 2001): Relevance Vector Regression (RVR) is an extension of SVR embedded in a Bayesian framework. Its characteristic feature is

that it imposes an explicit zero-mean Gaussian prior on the model parameters leading to a vector of independent hyperparameters that reduces the dataset. The behavior of the RVR is controlled by the type of kernel, which has to be chosen, while all other parameters are automatically estimated by the learning procedure itself. Here we used a linear, polynomial, or RBF kernel implemented with functions “rvmLinear” “rvmPoly” and “rvmRadial” respectively. The latter two kernels require tuning for scale and degree (polynomial) and for sigma (RBF).

(14)**Bayesian Regularized Neural Networks** (Perez-Rodriguez et al., 2013): This is a version of the feedforward artificial neural network (ANN) architecture, in which robustness is improved through Bayesian regularization of the ANN parameters. The model includes two layers: the input layer - consisting of independent variables - and the hidden layer of *S* number of neurons. Models were implemented using the “brnn” function and tuned for the number of neurons.

(15)**Extreme Gradient Boosting** (Chen and Guestrin, 2016): Extreme Gradient Boosting (XGBoost) is an ensemble decision-tree based gradient boosting algorithm that allows for modeling complex nonlinear relationships and interactions. The algorithm optimizes model performance through parallel (simultaneous) processing, regularization, tree pruning, optimal split (through a weighted quantile sketch algorithm), automatic missing data handling and built-in cross-validation. Tuning parameters involved the number of boosting iterations; maximum tree depth; eta (shrinkage parameter); gamma (minimum loss reduction); subsample ratio of columns; minimum sum of instance weights, and column subsample percentage. Models were implemented using “xgbTree” function.

As there are alternative functions for some of the algorithms examined here, in the results we refer to each algorithm using the name of the specific function used for implementation.

Computational efficiency for each algorithm was assessed by recording the total Central Processing Unit (CPU) time, and the average and maximum memory usage. All models were run on the ISMMS high-performance computing cluster.

Several analytic steps were common to all algorithms. As there are known sex differences in the rate of age-related changes (Brouwer et al., 2021; Wierenga et al., 2019; Wierenga et al., 2020), models were separately trained for males and females. Hyperparameter tuning (when required) was performed in the combined training set (n=2,105), using a grid search in a five-fold cross-validation scheme across five repeats. In each cross-validation 80% of the training sample was used to train the model and 20% was used to test the model parameters. Subsequently, the model was re-trained on the whole training dataset using the optimal hyperparameters identified through cross-validation. Finally, the generalizability of the model was tested in a hold-out test dataset (ABCD n=4,078).

The primary accuracy measure for each algorithm was the Mean Absolute Error (MAE) which represents the absolute difference between the neuroimaging-predicted age and the chronological age. For each algorithm, the abbreviation MAE_T refers to values obtained in the hold-out test dataset and MAE_{cv} refers to the mean cross-validation value in the training dataset. We also report two other commonly used accuracy measures: the Root Mean Square Error (RMSE), which is the standard deviation of the prediction errors, and the correlation between predicted and actual age. This correlation coefficient was not calculated in the ABCD data because of the narrow age range (less than 2 years). Based on these criteria we identified the three best performing algorithms which we evaluate further In the subsequent sections.

2.4 Calculating BrainAGE and corrected BrainAGE for the three best performing algorithms

BrainAGE in each individual was calculated by subtracting the chronological age from the age predicted by each of the three best performing algorithms. A positive BrainAGE value

indicates an older than expected brain-age for the given chronological age, and the opposite is the case for negative BrainAGE values. BrainAGE is typically overestimated in younger individuals and underestimated in older individuals. To counter this bias, multiple methods have been proposed (Beheshti et al., 2019; Cole et al., 2018). Here we used a robust approach introduced by Beheshti et al. (Beheshti et al., 2019), which relies on the slope (α) and intercept (β) of a linear regression model of BrainAGE against chronological age in the training set. This way an offset is calculated (as αΩ+β) and then subtracted from the estimated brain-age to yield a bias-free brain-age-gap-estimation (Beheshti et al., 2019), hereafter referred to as "corrected BrainAGE" ($BrainAGE_{corr}$).

2.5 Quantifying feature importance for age prediction in the three best performing algorithms

Estimates of the contribution of individual neuroimaging features to age prediction in each of the three best performing algorithms were obtained using Shapley Values (SV) implemented via the *fastshap* package Version 0.0.5 (Greenwell and Greenwell, 2020) in R. The SV derives from the cooperative game theory (Lundberg and Lee, 2017) where it is used to determine the contribution (payout) of each player to the total surplus generated in a coalition game. SVs accommodate non-linearity and have properties that makes their interpretation intuitive. For example, the sum of all the SVs of a model is equal to the accuracy of the model and features with the same SV contribute equally to the model.

2.6 Sensitivity and supplemental analyses in the three best performing algorithms

To test the effect of sex on model generalizability, we applied the parameters trained on one sex to the other and compared differences in BrainAGE using a two-sample Student's t test. Sensitivity analyses focused on the parcellation scheme, number of input features, sample size and number of repeats and cross-validation folds. Accordingly, we repeated the

analyses using features from (a) the Desikan-Killiany (DK) atlas (n of features=136); (b) the DK and subcortical Aseg atlas in FreeSurfer combined (n of features=157); (c) the 400-parcel Schaefer atlas with Aseg atlas (n of features=821); and (d) the 1000-parcel Schaefer atlas (n of features 2000). To test effect of sample size, the training dataset was randomly resampled with replacement in increments of 100, from 100-1500 (20 times each). Finally, we conducted the same analyses using 10 repeats and 10 cross-validation folds.

3. Results

3.1 Algorithm performance for age prediction

Linear algorithms, with the exception of Elastic Net Regression, performed less well (Table 1). Based on the MAE_T (Figure 1, Table 1), the top algorithms for both males and females were XGBoost (male MAE_T =1.57 years; female MAE_T =1.25 years), RF regression (male MAE_T =1.65 years; female MAE_T=, 1.23 years) and SVR with RBF kernel (MAE_T = 1.72 years; female MAE_T=,1.47 years) (Table 1). When corrected for age bias, the MAE_T of these algorithms reduced by 0.4-0.6 years in males and by 0.07-0.23 years in females (Figure 1). The correlation between age as predicted from the different models ranged between 0.34 to 0.99 for males and between 0.24 to 0.99 for females (Figure 2).

3.2 Computational speed and memory usage of each algorithm

The highest maximum memory usage was observed while training the Bayesian regularized neural networks, the boosted generalized additive model, and XGBoost algorithms (in that order). Highest average memory usage was seen with Bayesian regularized neural networks, the boosted generalized additive model, and the quantile regression with LASSO penalty. Bayesian regularized neural networks, XGBoost, and SVR with a polynomial kernel engaged CPU for the longest time. The generalized linear model, Kernel-partial least squares and principal component regression were the fastest algorithms with the lowest memory usage. Among the algorithms that performed best nonlinear kernelized versions had a favorable memory-computational speed profile (Supplemental Table 2).

3.3 BrainAGE and $BrainAGE_{corr}$ derived from the three best performing algorithms

We focus here only on the three best performing algorithms (i.e., XGBoost, RF regression and SVR with the RBF kernel) while information on all the models is presented in Supplemental Figures 3-5. The values presented refer to the models' performance in whole training sample using the optimized model parameters. In males, the mean (SD) values for BrainAGE were 1.33 (1.53) years for XGBoost, 1.55 (1.46) years for RF regression and 1.44

(1.57) years for SVR with the RBF kernel. In females, the corresponding values were 0.73 (1.52) years for XGBoost, 0.97 (1.34) years for RF regression and 1.04 (1.57) years for SVR with RBF kernel. In males, the mean (SD) values for $BrainAGE_{corr}$ were 0.22 (1.50) years for XGBoost, -0.03 (1.38) years for RF regression, and 0.56 (1.55) years for SVR with RBF kernel. In females, the corresponding values were -0.15 (1.49) years for XGBoost, -0.37 (1.27) years for RF regression, and 0.21 (1.55) years for SVR with the RBF kernel. As expected, there was a significantly negative correlation between BrainAGE and chronological age, which was mitigated after applying bias correction (Figure 2, Supplemental Figures 3-5).

3.4 Feature importance in brain-age prediction in the three best performing algorithms

Feature importance, as inferred by their SVs, varied considerably across models specified with XGBoost, RF regression, and SVR with the RBF kernel (Figure 4, Supplemental Figure 6, Supplemental Table 3). The values presented refer to the models' performance in whole training sample using the optimized model parameters. In RF regression, a few regions made very large contributions, with minimal contributions from other regions. In SVR with the RBF kernel, most features contributed to the model although the contribution of each feature was small. The profile of feature contributions in XGBoost was intermediate between the other two algorithms.

3.5 Sensitivity and supplemental analyses for the three best performing algorithms

Sex: Application of parameters from models trained on males to the entire sample, yielded marginally (but not significantly) higher BrainAGE values for females than males (maximum difference across models=0.2 years) (Supplemental Table 4) Similarly, application of parameters from models trained on females to the entire sample yield marginally (but nonsignificantly) higher BrainAGE for females than males (maximum difference across models=0.1 years) (Supplemental Table 4).

Parcellation scheme and number of input features: The different parcellation schemes had minimal influence on the MAE in any of the best performing algorithms. The number of features in the range examined (136-2,000) had minimal impact on MAE and notably the Schaefer-1000 parcellation did not outperform the Schaefer-400 parcellation used in the main analyses (Figure 4A). The same pattern was seen in females and males (Supplemental Figure 7).

Sample size**:** MAE_T increased in line with sample size increase up to a size of 500 participants and it plateaued thereafter. The corrected $MAE_{T,}$ on the other hand, was more stable with limited change across sample size (Figure 5B, Supplemental Figure 8).

Number of cross-validation folds and repeats: In the main analyses, we used 5-fold repeats and 5-fold cross-validations. Using 10 instead of 5 folds and repeats did not improve performance and in the case of XGBoost, we noted markedly worse performance as evidenced by an increase of about 3 years in the MAE_T (Supplemental Table 5).

4. Discussion

In the present study, we undertook a comprehensive comparison of machine learning algorithms for quantifying $BrainAGE_{corr}$ as a proxy for biological brain age during development. We identified three algorithms, namely XGBoost, RF regression, and SVR with the RBF kernel, that outperformed all others in terms of accuracy while being computationally efficient. Notably we also show that brain-based age prediction was suboptimal in models using linear algorithms.

As predicted XGBoost and RF regression, which are both ensemble tree-based algorithms, performed well in terms of their accuracy and generalizability to unseen samples. A decision tree is a machine learning algorithm that partitions the data into subsets based on conditional statements. Although each tree has generally low predictive performance, their combination (ensemble) improves generalizability without sacrificing accuracy (Qi, 2012). Further advantages of these methods, particularly in the context of neuroimaging datasets, is that they are nonparametric, they do not involve assumptions about the distribution for the data and can account for nonlinear effects and interactions, which may be particularly relevant in modeling developmental brain-age. They also require minimal or no preparation or cleaning of the input sMRI features as they can handle multicollinear data without losing accuracy. Notably, these algorithms were relatively insensitive to the number of the neuroimaging features, when the size of the feature sets ranged between 136 to 2000, with the mid-point feature set (n=841) being relatively better.

Similar to prior studies (Kaufmann et al., 2019; Lee et al., 2021), we demonstrated that linear algorithms consistently underperformed compared to other algorithms. This may be due to the multicollinearity of the neuroimaging, as suggested by the fact that linear algorithms that are based on covariance (such as SPLS regression or PCA regression) performed slightly better. Further, the general underperformance of linear models may also reflect the fact that

they do not account for nonlinear and interactive associations between brain imaging features and age.

A major concern in neuroimaging research is the effect of site on the generalizability of machine learning models (Dockes et al., 2021; Solanes et al., 2021). Sites may differ in terms of scanner infrastructure, acquisition protocols and neuroimaging feature extraction pipelines as well as sample composition. Here the post-acquisition extraction of neuroimaging features was undertaken for all cohorts using the same pipeline which may have reduced variability in the neuroimaging feature set. However, all other parameters differed between cohorts (and between recruitment sites within cohorts). Yet we showed that the best performing algorithms showed excellent generalizability to the hold-out datasets, which is likely to reflect the robustness of these algorithms. Additionally, the inclusion of observations from multiple sites in the training dataset may have forced the ML algorithms to select and weight features that are robust to site differences, therefore reducing the dependence of the model on the effects of site.

We observed a lower MAE for females compared to males across most models. This has been reported in prior studies (Brouwer et al., 2021; Wierenga et al., 2019; Wierenga et al., 2020), and can be attributed to either biological differences, i.e., female brain showing less variability or confounding, i.e., males may move more, on average, than females which could make their brain measurements less accurate. Brouwer and colleagues (Brouwer et al., 2021) demonstrated that in individuals ages 9-23, females have higher sMRI-derived BrainAGE than their male counterparts. The same pattern was reported by Tu and colleagues (Tu et al., 2019) using sMRI data from 118 males and 147 females, aged 5-18, from the NIH MRI Study of Normal Brain Development. In these studies, as well as in ours, sex differences in BrainAGE are small and within the range of the MAE for brain-predicted age.

We acknowledge that the list of algorithms evaluated is not exhaustive but provides a good coverage of the many models that are currently available. We were unable to account for potential influences of race and ethnicity as such information was either absent or not uniformly coded in the cohorts used for model training. Based on the racial constitution of the general population, in the countries of the recruitment sites, we anticipate an over-representation of white individuals. As more data becomes available on other racial/ethnic groups, it should be possible to address this issue in future studies.

In summary, using a wide range of ML algorithms on geographically diverse datasets of young people, we showed that tree-based followed by nonlinear kernel-based algorithms offer robust, accurate, and generalizable solution for predicting age based on brain morphological features. Findings of the present study can be used as a guide for quantifying brain maturation during development and its contribution to functional and behavioral outcomes.

Acknowledgements: Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093, and U01DA041025. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/scientists/workgroups/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. Dr S.F. received support from the National Institute of Mental Health (R01-MH113619).

This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Author statement

AM: Conceptualization, Methodology, Investigation, Data Curation, Formal analysis, Writing - Original Draft, Writing - Review & Editing

SF: Supervision, Conceptualization, Writing - Original Draft, Writing - Review & Editing

RSK: Supervision, Conceptualization, Writing - Review & Editing

HCW: Supervision, Conceptualization, Review & Editing

DCG: Supervision - Review & Editing

PMT: Supervision - Review & Editing

Data and Code Availability: All data used in this study are publicly available and can be accessed through following websites:
ABCD: https://nda.nih.gov/abcd;
ABIDE: http://fcon_1000.projects.nitrc.org/indi/abide/;
ABIDE 2: http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html;
ADHD-200: http://fcon_1000.projects.nitrc.org/indi/adhd200/# ;
CMI/HBN: http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/;
HCP-D: https://www.humanconnectome.org/study/hcp-lifespan-development .
Analysis code is available at https://github.com/AmirhosseinModabbernia/ABCD

References

Alexander, L.M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., Chan, E., Citera, D., Clucas, J., Cohen, S., Dufek, S., Eaves, M., Fradera, B., Gardner, J., Grant-Villegas, N., Green, G., Gregory, C., Hart, E., Harris, S., Horton, M., Kahn, D., Kabotyanski, K., Karmel, B., Kelly, S.P., Kleinman, K., Koo, B., Kramer, E., Lennon, E., Lord, C., Mantello, G., Margolis, A., Merikangas, K.R., Milham, J., Minniti, G., Neuhaus, R., Levine, A., Osman, Y., Parra, L.C., Pugh, K.R., Racanello, A., Restrepo, A., Saltzman, T., Septimus, B., Tobe, R., Waltz, R., Williams, A., Yeo, A., Castellanos, F.X., Klein, A., Paus, T., Leventhal, B.L., Craddock, R.C., Koplewicz, H.S., Milham, M.P., 2017. An open resource for transdiagnostic research in pediatric mental health and learning disorders. Sci Data 4, 170181.

Anaturk, M., Kaufmann, T., Cole, J.H., Suri, S., Griffanti, L., Zsoldos, E., Filippini, N., Singh-Manoux, A., Kivimaki, M., Westlye, L.T., Ebmeier, K.P., de Lange, A.G., 2021. Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging. Hum Brain Mapp 42, 1626-1640.

Awad, M., Khanna, R., 2015. Support vector regression. Efficient learning machines. Springer, pp. 67-80.

Ball, G., Kelly, C.E., Beare, R., Seal, M.L., 2021. Individual variation underlying brain age estimates in typical development. Neuroimage 235, 118036.

Beheshti, I., Nugent, S., Potvin, O., Duchesne, S., 2019. Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. Neuroimage Clin 24, 102063.

Boyle, R., Jollans, L., Rueda-Delgado, L.M., Rizzo, R., Yener, G.G., McMorrow, J.P., Knight, S.P., Carey, D., Robertson, I.H., Emek-Savas, D.D., Stern, Y., Kenny, R.A., Whelan, R., 2021. Brain-predicted age difference score is related to specific cognitive functions: a multi-site replication analysis. Brain Imaging Behav 15, 327-345.

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Brouwer, R.M., Schutte, J., Janssen, R., Boomsma, D.I., Hulshoff Pol, H.E., Schnack, H.G., 2021. The Speed of Development of Adolescent Brain Age Depends on Sex and Is Genetically Determined. Cereb Cortex 31, 1296-1306.

Bryce, N.V., Flournoy, J.C., Guassi Moreira, J.F., Rosen, M.L., Sambook, K.A., Mair, P., McLaughlin, K.A., 2021. Brain parcellation selection: An overlooked decision point with meaningful effects on individual differences in resting-state functional connectivity. Neuroimage 243, 118487.

Bühlmann, P., Yu, B., 2003. Boosting with the L 2 loss: regression and classification. Journal of the American Statistical Association 98, 324-339.

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

Chun, H., Keleş, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the royal statistical society: series B (statistical methodology) 72, 3-25.

Chung, Y., Addington, J., Bearden, C.E., Cadenhead, K., Cornblatt, B., Mathalon, D.H., McGlashan, T., Perkins, D., Seidman, L.J., Tsuang, M., Walker, E., Woods, S.W., McEwen, S., van Erp, T.G.M., Cannon, T.D., 2018. Use of Machine Learning to Determine Deviance in Neuroanatomical Maturity Associated With Future Psychosis in Youths at Clinically High Risk. JAMA Psychiatry 75, 960-968.

Cole, J.H., Ritchie, S.J., Bastin, M.E., Valdes Hernandez, M.C., Munoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., Wray, N.R., Redmond, P., Marioni, R.E., Starr, J.M., Cox, S.R., Wardlaw, J.M., Sharp, D.J., Deary, I.J., 2018. Brain age predicts mortality. Mol Psychiatry 23, 1385-1392.

Consortium, H.D., 2012. The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience. Front Syst Neurosci 6, 62.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273-297.

Dayal, B.S., MacGregor, J.F., 1997. Improved PLS algorithms. Journal of Chemometrics: A Journal of the Chemometrics Society 11, 73-85.
Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L.M., Bookheimer, S.Y., Braden, B.B., Byrge, L., Castellanos, F.X., Dapretto, M., Delorme, R., Fair, D.A., Fishman, I., Fitzgerald, J., Gallagher, L., Keehn, R.J., Kennedy, D.P., Lainhart, J.E., Luna, B., Mostofsky, S.H., Muller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Solomon, M., Toro, R., Vaidya, C.J., Wenderoth, N., White, T., Craddock, R.C., Lord, C., Leventhal, B., Milham, M.P., 2017. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data 4, 170010.
Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Muller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19, 659-667.
Dima, D., Modabbernia, A., Papachristou, E., Doucet, G.E., Agartz, I., Aghajani, M., Akudjedu, T.N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K.I., Andersson, M., Andreasen, N.C., Andreassen, O.A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., Bonvino, A., Boomsma, D.I., Borgwardt, S., Bourque, J., Brandeis, D., Breier, A., Brodaty, H., Brouwer, R.M., Buitelaar, J.K., Busatto, G.F., Buckner, R.L., Calhoun, V., Canales-Rodriguez, E.J., Cannon, D.M., Caseras, X., Castellanos, F.X., Cervenka, S., Chaim-Avancini, T.M., Ching, C.R.K., Chubar, V., Clark, V.P., Conrod, P., Conzelmann, A., Crespo-Facorro, B., Crivello, F., Crone, E.A., Dannlowski, U., Dale, A.M., Davey, C., de Geus, E.J.C., de Haan, L., de Zubicaray, G.I., den Braber, A., Dickie, E.W., Di Giorgio, A., Doan, N.T., Dorum, E.S., Ehrlich, S., Erk, S., Espeseth, T., Fatouros-Bergman, H., Fisher, S.E., Fouche, J.P., Franke, B., Frodl, T., Fuentes-Claramonte, P., Glahn, D.C., Gotlib, I.H., Grabe, H.J., Grimm, O., Groenewold, N.A., Grotegerd, D., Gruber, O., Gruner, P., Gur, R.E., Gur, R.C., Hahn, T., Harrison, B.J., Hartman, C.A., Hatton, S.N., Heinz, A., Heslenfeld, D.J., Hibar, D.P., Hickie, I.B., Ho, B.C., Hoekstra, P.J., Hohmann, S., Holmes, A.J., Hoogman, M., Hosten, N., Howells, F.M., Hulshoff Pol, H.E., Huyser, C., Jahanshad, N., James, A., Jernigan, T.L., Jiang, J., Jonsson, E.G., Joska, J.A., Kahn, R., Kalnin, A., Kanai, R., Klein, M., Klyushnik, T.P., Koenders, L., Koops, S., Kramer, B., Kuntsi, J., Lagopoulos, J., Lazaro, L., Lebedeva, I., Lee, W.H., Lesch, K.P., Lochner, C., Machielsen, M.W.J., Maingault, S., Martin, N.G., Martinez-Zalacain, I., Mataix-Cols, D., Mazoyer, B., McDonald, C., McDonald, B.C., McIntosh, A.M., McMahon, K.L., McPhilemy, G., Meinert, S., Menchon, J.M., Medland, S.E., Meyer-Lindenberg, A., Naaijen, J., Najt, P., Nakao, T., Nordvik, J.E., Nyberg, L., Oosterlaan, J., de la Foz, V.O., Paloyelis, Y., Pauli, P., Pergola, G., Pomarol-Clotet, E., Portella, M.J., Potkin, S.G., Radua, J., Reif, A., Rinker, D.A., Roffman, J.L., Rosa, P.G.P., Sacchet, M.D., Sachdev, P.S., Salvador, R., Sanchez-Juan, P., Sarro, S., Satterthwaite, T.D., Saykin, A.J., Serpa, M.H., Schmaal, L., Schnell, K., Schumann, G., Sim, K., Smoller, J.W., Sommer, I., Soriano-Mas, C., Stein, D.J., Strike, L.T., Swagerman, S.C., Tamnes, C.K., Temmingh, H.S., Thomopoulos, S.I., Tomyshev, A.S., Tordesillas-Gutierrez, D., Trollor, J.N., Turner, J.A., Uhlmann, A., van den Heuvel, O.A., van den Meer, D., van der Wee, N.J.A., van Haren, N.E.M., Van't Ent, D., van Erp, T.G.M., Veer, I.M., Veltman, D.J., Voineskos, A., Volzke, H., Walter, H., Walton, E., Wang, L., Wang, Y., Wassink, T.H., Weber, B., Wen, W., West, J.D., Westlye, L.T., Whalley, H., Wierenga, L.M., Williams, S.C.R., Wittfeld, K., Wolf, D.H., Worker, A., Wright, M.J., Yang, K., Yoncheva, Y., Zanetti, M.V., Ziegler, G.C., Thompson, P.M., Frangou, S., Karolinska Schizophrenia, P., 2021. Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3-90 years. Hum Brain Mapp.
Dockes, J., Varoquaux, G., Poline, J.B., 2021. Preventing dataset shift from breaking machine-learning biomarkers. Gigascience 10.

Elliott, M.L., Belsky, D.W., Knodt, A.R., Ireland, D., Melzer, T.R., Poulton, R., Ramrakha, S., Caspi, A., Moffitt, T.E., Hariri, A.R., 2019. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. Mol Psychiatry.
Erus, G., Battapady, H., Satterthwaite, T.D., Hakonarson, H., Gur, R.E., Davatzikos, C., Gur, R.C., 2015. Imaging patterns of brain development and their relationship to cognition. Cereb Cortex 25, 1676-1684.
Fjell, A.M., Westlye, L.T., Amlien, I., Tamnes, C.K., Grydeland, H., Engvig, A., Espeseth, T., Reinvang, I., Lundervold, A.J., Lundervold, A., Walhovd, K.B., 2015. High-expanding cortical regions in human development and evolution are related to higher intellectual abilities. Cereb Cortex 25, 26-34.
Frangou, S., Modabbernia, A., Williams, S.C.R., Papachristou, E., Doucet, G.E., Agartz, I., Aghajani, M., Akudjedu, T.N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K.I., Andersson, M., Andreasen, N.C., Andreassen, O.A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., Bonvino, A., Boomsma, D.I., Borgwardt, S., Bourque, J., Brandeis, D., Breier, A., Brodaty, H., Brouwer, R.M., Buitelaar, J.K., Busatto, G.F., Buckner, R.L., Calhoun, V., Canales-Rodriguez, E.J., Cannon, D.M., Caseras, X., Castellanos, F.X., Cervenka, S., Chaim-Avancini, T.M., Ching, C.R.K., Chubar, V., Clark, V.P., Conrod, P., Conzelmann, A., Crespo-Facorro, B., Crivello, F., Crone, E.A., Dale, A.M., Dannlowski, U., Davey, C., de Geus, E.J.C., de Haan, L., de Zubicaray, G.I., den Braber, A., Dickie, E.W., Di Giorgio, A., Doan, N.T., Dorum, E.S., Ehrlich, S., Erk, S., Espeseth, T., Fatouros-Bergman, H., Fisher, S.E., Fouche, J.P., Franke, B., Frodl, T., Fuentes-Claramonte, P., Glahn, D.C., Gotlib, I.H., Grabe, H.J., Grimm, O., Groenewold, N.A., Grotegerd, D., Gruber, O., Gruner, P., Gur, R.E., Gur, R.C., Hahn, T., Harrison, B.J., Hartman, C.A., Hatton, S.N., Heinz, A., Heslenfeld, D.J., Hibar, D.P., Hickie, I.B., Ho, B.C., Hoekstra, P.J., Hohmann, S., Holmes, A.J., Hoogman, M., Hosten, N., Howells, F.M., Hulshoff Pol, H.E., Huyser, C., Jahanshad, N., James, A., Jernigan, T.L., Jiang, J., Jonsson, E.G., Joska, J.A., Kahn, R., Kalnin, A., Kanai, R., Klein, M., Klyushnik, T.P., Koenders, L., Koops, S., Kramer, B., Kuntsi, J., Lagopoulos, J., Lazaro, L., Lebedeva, I., Lee, W.H., Lesch, K.P., Lochner, C., Machielsen, M.W.J., Maingault, S., Martin, N.G., Martinez-Zalacain, I., Mataix-Cols, D., Mazoyer, B., McDonald, C., McDonald, B.C., McIntosh, A.M., McMahon, K.L., McPhilemy, G., Meinert, S., Menchon, J.M., Medland, S.E., Meyer-Lindenberg, A., Naaijen, J., Najt, P., Nakao, T., Nordvik, J.E., Nyberg, L., Oosterlaan, J., de la Foz, V.O., Paloyelis, Y., Pauli, P., Pergola, G., Pomarol-Clotet, E., Portella, M.J., Potkin, S.G., Radua, J., Reif, A., Rinker, D.A., Roffman, J.L., Rosa, P.G.P., Sacchet, M.D., Sachdev, P.S., Salvador, R., Sanchez-Juan, P., Sarro, S., Satterthwaite, T.D., Saykin, A.J., Serpa, M.H., Schmaal, L., Schnell, K., Schumann, G., Sim, K., Smoller, J.W., Sommer, I., Soriano-Mas, C., Stein, D.J., Strike, L.T., Swagerman, S.C., Tamnes, C.K., Temmingh, H.S., Thomopoulos, S.I., Tomyshev, A.S., Tordesillas-Gutierrez, D., Trollor, J.N., Turner, J.A., Uhlmann, A., van den Heuvel, O.A., van den Meer, D., van der Wee, N.J.A., van Haren, N.E.M., van 't Ent, D., van Erp, T.G.M., Veer, I.M., Veltman, D.J., Voineskos, A., Volzke, H., Walter, H., Walton, E., Wang, L., Wang, Y., Wassink, T.H., Weber, B., Wen, W., West, J.D., Westlye, L.T., Whalley, H., Wierenga, L.M., Wittfeld, K., Wolf, D.H., Worker, A., Wright, M.J., Yang, K., Yoncheva, Y., Zanetti, M.V., Ziegler, G.C., Karolinska Schizophrenia, P., Thompson, P.M., Dima, D., 2021. Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3-90 years. Hum Brain Mapp.
Franke, K., Gaser, C., 2019. Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? Front Neurol 10, 789.
Franke, K., Luders, E., May, A., Wilke, M., Gaser, C., 2012. Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. Neuroimage 63, 1305-1312.
Franke, K., Ziegler, G., Kloppel, S., Gaser, C., Alzheimer's Disease Neuroimaging, I., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50, 883-892.
Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R.Z., Heeringa, S., Jernigan, T., Potter, A., Thompson, W., Zahs, D., 2018. Recruiting the ABCD sample: Design considerations and procedures. Dev Cogn Neurosci 32, 16-22.

Gaser, C., Franke, K., Kloppel, S., Koutsouleris, N., Sauer, H., Alzheimer's Disease Neuroimaging, I., 2013. BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. PLoS One 8, e67346.
Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S., 2008. A weakly informative default prior distribution for logistic and other regression models. The annals of applied statistics 2, 1360-1383.
Greenwell, B., Greenwell, M.B., 2020. Package 'fastshap'.
Harms, M.P., Somerville, L.H., Ances, B.M., Andersson, J., Barch, D.M., Bastiani, M., Bookheimer, S.Y., Brown, T.B., Buckner, R.L., Burgess, G.C., Coalson, T.S., Chappell, M.A., Dapretto, M., Douaud, G., Fischl, B., Glasser, M.F., Greve, D.N., Hodge, C., Jamison, K.W., Jbabdi, S., Kandala, S., Li, X., Mair, R.W., Mangia, S., Marcus, D., Mascali, D., Moeller, S., Nichols, T.E., Robinson, E.C., Salat, D.H., Smith, S.M., Sotiropoulos, S.N., Terpstra, M., Thomas, K.M., Tisdall, M.D., Ugurbil, K., van der Kouwe, A., Woods, R.P., Zollei, L., Van Essen, D.C., Yacoub, E., 2018. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. Neuroimage 183, 972-984.
Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnaes, D., Barch, D.M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M.K., Boen, E., Borgwardt, S., Brandt, C.L., Buitelaar, J., Celius, E.G., Cervenka, S., Conzelmann, A., Cordova-Palomera, A., Dale, A.M., de Quervain, D.J.F., Di Carlo, P., Djurovic, S., Dorum, E.S., Eisenacher, S., Elvsashagen, T., Espeseth, T., Fatouros-Bergman, H., Flyckt, L., Franke, B., Frei, O., Haatveit, B., Haberg, A.K., Harbo, H.F., Hartman, C.A., Heslenfeld, D., Hoekstra, P.J., Hogestol, E.A., Jernigan, T.L., Jonassen, R., Jonsson, E.G., Karolinska Schizophrenia, P., Kirsch, P., Kloszewska, I., Kolskar, K.K., Landro, N.I., Le Hellard, S., Lesch, K.P., Lovestone, S., Lundervold, A., Lundervold, A.J., Maglanoc, L.A., Malt, U.F., Mecocci, P., Melle, I., Meyer-Lindenberg, A., Moberget, T., Norbom, L.B., Nordvik, J.E., Nyberg, L., Oosterlaan, J., Papalino, M., Papassotiropoulos, A., Pauli, P., Pergola, G., Persson, K., Richard, G., Rokicki, J., Sanders, A.M., Selbaek, G., Shadrin, A.A., Smeland, O.B., Soininen, H., Sowa, P., Steen, V.M., Tsolaki, M., Ulrichsen, K.M., Vellas, B., Wang, L., Westman, E., Ziegler, G.C., Zink, M., Andreassen, O.A., Westlye, L.T., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat Neurosci 22, 1617-1623.
Lee, W.H., Antoniades, M., Schnack, H.G., Kahn, R.S., Frangou, S., 2021. Brain age prediction in schizophrenia: Does the choice of machine learning algorithm matter? Psychiatry Res Neuroimaging 310, 111270.
Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M.L., Witte, A.V., Villringer, A., Margulies, D.S., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage 148, 179-188.
Luna, A., Bernanke, J., Kim, K., Aw, N., Dworkin, J.D., Cha, J., Posner, J., 2021. Maturity of gray matter structures and white matter connectomes, and their relationship with psychiatric symptoms in youth. Hum Brain Mapp.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems, pp. 4768-4777.
Mills, K.L., Siegmund, K.D., Tamnes, C.K., Ferschmann, L., Wierenga, L.M., Bos, M.G.N., Luna, B., Li, C., Herting, M.M., 2021. Inter-individual variability in structural brain development from late childhood to young adulthood. Neuroimage 242, 118450.
Paus, T., 2005. Mapping brain maturation and cognitive development during adolescence. Trends Cogn Sci 9, 60-68.
Perez-Rodriguez, P., Gianola, D., Weigel, K.A., Rosa, G.J., Crossa, J., 2013. Technical note: An R package for fitting Bayesian regularized neural networks with applications in animal breeding. J Anim Sci 91, 3522-3531.
Qi, Y., 2012. Random forest for bioinformatics. Ensemble machine learning. Springer, pp. 307-323.

Raznahan, A., Shaw, P.W., Lerch, J.P., Clasen, L.S., Greenstein, D., Berman, R., Pipitone, J., Chakravarty, M.M., Giedd, J.N., 2014. Longitudinal four-dimensional mapping of subcortical anatomy in human development. Proc Natl Acad Sci U S A 111, 1592-1597.
Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb Cortex 28, 3095-3114.
Shao, X., Wang, W., Hou, Z., Cai, W., 2006. A new regression method based on independent component analysis. Talanta 69, 676-680.
Solanes, A., Palau, P., Fortea, L., Salvador, R., González-Navarro, L., Llach, C.D., Valentí, M., Vieta, E., Radua, J., 2021. Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. Psychiatry Res Neuroimaging 314, 111313.
Spear, L.P., 2000. The adolescent brain and age-related behavioral manifestations. Neurosci Biobehav Rev 24, 417-463.
Tamnes, C.K., Herting, M.M., Goddings, A.L., Meuwese, R., Blakemore, S.J., Dahl, R.E., Guroglu, B., Raznahan, A., Sowell, E.R., Crone, E.A., Mills, K.L., 2017. Development of the Cerebral Cortex across Adolescence: A Multisample Study of Inter-Related Longitudinal Changes in Cortical Volume, Surface Area, and Thickness. J Neurosci 37, 3402-3412.
Tau, G.Z., Peterson, B.S., 2010. Normal development of brain circuits. Neuropsychopharmacology 35, 147-168.
Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research 1, 211-244.
Tu, Y., Fu, Z., Maleki, N., 2019. When does the youthfulness of the female brain emerge? Proc Natl Acad Sci U S A 116, 10632-10633.
Valizadeh, S.A., Hanggi, J., Merillat, S., Jancke, L., 2017. Age prediction on the basis of brain anatomical measures. Hum Brain Mapp 38, 997-1008.
Valk, S.L., Xu, T., Margulies, D.S., Masouleh, S.K., Paquola, C., Goulas, A., Kochunov, P., Smallwood, J., Yeo, B.T.T., Bernhardt, B.C., Eickhoff, S.B., 2020. Shaping brain structure: Genetic and phylogenetic axes of macroscale organization of cortical thickness. Sci Adv 6.
Wierenga, L.M., Bos, M.G.N., van Rossenberg, F., Crone, E.A., 2019. Sex Effects on Development of Brain Structure and Executive Functions: Greater Variance than Mean Effects. J Cogn Neurosci 31, 730-753.
Wierenga, L.M., Doucet, G.E., Dima, D., Agartz, I., Aghajani, M., Akudjedu, T.N., Albajes-Eizagirre, A., Alnaes, D., Alpert, K.I., Andreassen, O.A., Anticevic, A., Asherson, P., Banaschewski, T., Bargallo, N., Baumeister, S., Baur-Streubel, R., Bertolino, A., Bonvino, A., Boomsma, D.I., Borgwardt, S., Bourque, J., den Braber, A., Brandeis, D., Breier, A., Brodaty, H., Brouwer, R.M., Buitelaar, J.K., Busatto, G.F., Calhoun, V.D., Canales-Rodríguez, E.J., Cannon, D.M., Caseras, X., Castellanos, F.X., Chaim-Avancini, T.M., Ching, C.R., Clark, V.P., Conrod, P.J., Conzelmann, A., Crivello, F., Davey, C.G., Dickie, E.W., Ehrlich, S., Van't Ent, D., Fisher, S.E., Fouche, J.P., Franke, B., Fuentes-Claramonte, P., de Geus, E.J., Di Giorgio, A., Glahn, D.C., Gotlib, I.H., Grabe, H.J., Gruber, O., Gruner, P., Gur, R.E., Gur, R.C., Gurholt, T.P., de Haan, L., Haatveit, B., Harrison, B.J., Hartman, C.A., Hatton, S.N., Heslenfeld, D.J., van den Heuvel, O.A., Hickie, I.B., Hoekstra, P.J., Hohmann, S., Holmes, A.J., Hoogman, M., Hosten, N., Howells, F.M., Hulshoff Pol, H.E., Huyser, C., Jahanshad, N., James, A.C., Jiang, J., Jönsson, E.G., Joska, J.A., Kalnin, A.J., Klein, M., Koenders, L., Kolskår, K.K., Krämer, B., Kuntsi, J., Lagopoulos, J., Lazaro, L., Lebedeva, I.S., Lee, P.H., Lochner, C., Machielsen, M.W., Maingault, S., Martin, N.G., Martínez-Zalacaín, I., Mataix-Cols, D., Mazoyer, B., McDonald, B.C., McDonald, C., McIntosh, A.M., McMahon, K.L., McPhilemy, G., van der Meer, D., Menchón, J.M., Naaijen, J., Nyberg, L., Oosterlaan, J., Paloyelis, Y., Pauli, P., Pergola, G., Pomarol-Clotet, E., Portella, M.J., Radua, J., Reif, A., Richard, G., Roffman, J.L., Rosa, P.G., Sacchet, M.D., Sachdev, P.S., Salvador, R., Sarró, S., Satterthwaite, T.D., Saykin, A.J., Serpa, M.H., Sim, K., Simmons, A., Smoller, J.W., Sommer, I.E., Soriano-Mas, C., Stein, D.J., Strike, L.T., Szeszko, P.R., Temmingh, H.S., Thomopoulos, S.I., Tomyshev, A.S., Trollor, J.N., Uhlmann, A., Veer, I.M., Veltman, D.J., Voineskos, A., Völzke, H., Walter, H., Wang, L., Wang, Y.,

Weber, B., Wen, W., West, J.D., Westlye, L.T., Whalley, H.C., Williams, S.C., Wittfeld, K., Wolf, D.H., Wright, M.J., Yoncheva, Y.N., Zanetti, M.V., Ziegler, G.C., de Zubicaray, G.I., Thompson, P.M., Crone, E.A., Frangou, S., Tamnes, C.K., 2020. Greater male than female variability in regional brain structure across the lifespan. Hum Brain Mapp.
Williams, C.K., Barber, D., 1998. Bayesian classification with Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 1342-1351.
Wu, Y., Liu, Y., 2009. Variable selection in quantile regression. Statistica Sinica, 801-817.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67, 301-320.

Table 1. Model Performance in the Hold-out Test and Cross-Validation Sets

	Hold-out Test-set performance			Cross-Validation performance		
Algorithm (function name in caret package)	**MAE_T**	**$RMSE_T$**	**Bias-Adjusted MAE**	**MAE_{cv}**	**$RMSE_{cv}$**	**Correlation**
Males						
Extreme Gradient Boosting (xgbTree)	1.57	2.03	1.16	1.7	2.18	0.83
Random Forest Regression (rf)	1.65	2.13	1.09	1.99	2.52	0.77
Support Vector Regression -Radial (svmRadial)	1.72	2.14	1.29	1.6	2.1	0.84
Support Vector Regression -Polynomial (svmPoly)	1.74	2.15	1.3	1.63	2.14	0.83
Relevance Vector Regression-Polynomial (rvmPoly)	1.78	2.2	1.38	1.65	2.16	0.83
Gaussian Processes Polynomial (gaussprPoly)	1.8	2.22	1.2	1.63	2.13	0.83
Gaussian Processes Radial (gaussprRadial)	1.81	2.22	1.34	1.71	2.18	0.83
Generalized Additive Model with Boosting (gamboost)	1.82	2.25	1.3	1.81	2.31	0.80
Sparse Partial Least Squares (spls)	1.86	2.29	1.44	1.77.	2.26	0.81
Kernel Partial Least Squares (kernelpls)	1.86	2.29	1.44	1.77	2.26	0.81
Elastic Net Regression (enet)	1.9	2.32	1.44	1.73	2.21	0.82
Quantile Regression with LASSO penalty (rqlasso)	1.91	2.33	1.42	1.77	2.26	0.81
Relevance Vector Regression-Radial (rvmRadial)	1.94	2.39	1.53	1.71	2.23	0.82
Bayesian Regularized Neural Network (brnn)	1.99	2.51	1.69	1.93	2.44	0.79
Independent Component Regression (icr)	2.1	2.55	1.37	2.19	2.72	0.71
Principal Component Regression (pcr)	2.1	2.55	1.37	2.19	2.72	0.71
Support Vector Regression-Linear (svmLinear3)	2.7	3.37	2.5	2.87	3.67	0.65
Gaussian Processes-	2.77	3.45	2.52	3.21	4.1	0.6

Table 1. Model Performance in the Hold-out Test and Cross-Validation Sets

	Hold-out Test-set performance			Cross-Validation performance		
Algorithm (function name in caret package)	**MAE_T**	**$RMSE_T$**	**Bias-Adjusted MAE**	**MAE_{cv}**	**$RMSE_{cv}$**	**Correlation**
Linear (gaussprLinear)						
Generalized Linear Model (glm)	2.81	3.49	2.56	3.43	4.39	0.57
Bayesian Generalized Linear Model (bayesglm)	2.81	3.5	2.56	3.43	4.39	0.57
Relevance Vector Machine-Linear (rvmRaidal)	11.09	11.22	1.32	12.89	13.14	0.76
Females						
Random Forest Regression (rf)	1.23	1.66	1.08	1.82	2.33	0.82
Extreme Gradient Boosting (xgbTree)	1.25	1.69	1.19	1.59	2.08	0.85
Support Vector Regression -Radial (svmRadial)	1.47	1.89	1.21	1.6	2.08	0.85
Support Vector Regression -Polynomial (svmPoly)	1.47	1.89	1.24	1.58	2.06	0.85
Gaussian Processes Polynomial (gaussprPoly)	1.48	1.9	1.2	1.61	2.09	0.85
Generalized Additive Model with Boosting (gamboost)	1.49	1.91	1.33	1.71	2.19	0.83
Relevance Vector Regression -Polynomial (rvmPoly)	1.51	1.94	1.28	1.63	2.13	0.84
Gaussian Processes Radial (gaussprRadial)	1.52	1.95	1.24	1.65	2.13	0.85
Relevance Vector Regression -Radial (rvmRadial)	1.64	2.08	1.44	1.75	2.22	0.83
Quantile Regression with LASSO penalty (rqlasso)	1.66	2.08	1.43	1.72	2.24	0.82
Independent Component Regression (icr)	1.67	2.11	1.44	1.93	2.42	0.79
Principal Component Regression (pcr)	1.67	2.11	1.44	1.93	2.43	0.79
Elastic Net Regression (enet)	1.7	2.15	1.51	1.7	2.18	0.84

Table 1. Model Performance in the Hold-out Test and Cross-Validation Sets

	Hold-out Test-set performance			Cross-Validation performance		
Algorithm (function name in caret package)	**MAE_T**	**$RMSE_T$**	**Bias-Adjusted MAE**	**MAE_{cv}**	**$RMSE_{cv}$**	**Correlation**
Kernel Partial Least Squares (kernelpls)	1.78	2.21	1.5	1.72	2.19	0.83
Sparse Partial Least Squares (spls)	1.78	2.21	1.5	1.71	2.18	0.83
Bayesian Regularized Neural Network (brnn)	1.97	2.52	1.79	1.95	2.44	0.79
Support Vector Regression-Linear (svmLinear3)	3.66	4.58	3.51	3.3	4.18	0.59
Gaussian Processes-Linear (gaussprLinear)	4.17	5.23	3.97	4	5.08	0.51
Bayesian Generalized Linear Model (bayesglm)	5.24	6.61	5.06	4.6	5.84	0.46
Generalized Linear Model (glm)	5.3	6.69	10.47	123.87	160.56	0.1
Relevance Vector Machine-Linear (rvmLinear)	11.67	11.83	1.47	12.71	12.95	0.78

LASSO: least absolute shrinkage and selection operator**;** MAE: Mean Absolute Error; RMSE: Root Mean Squared Error

Figures

Figure 1. Performance of the 21 tested machine learning algorithms. Figure demonstrates model performance in the hold-out test set in males and female. **A.** Absolute Error; **B.** Bias-Corrected Absolute Error. bayesglm: Bayesian Generalized Linear Model; brnn: Bayesian Regularized Neural Network; enet: Elastic Net Regression; gamboost: Generalized Additive Model with Boosting; gaussprLinear: Gaussian Processes Regression Linear; gaussprPoly: Gaussian Processes Regression Polynomial; gaussprRadial: Gaussian Processes Regression Radial; glm: Generalized Linear Model; icr: Independent Component Regression; kernelpls: Kernel Partial Least Squares; pcr: Principal Component Regression; rf: Random Forest; rqlasso: Quantile Regression with LASSO penalty; rvmLinear: Relevance Vector Machine-Linear; rvmPoly: Relevance Vector Machine-Polynomial; rvmRadial: Relevance Vector Machine-Radial; spls: Sparse Partial Least Squares; svmLinear3: Support Vector Regression-Linear; svmPoly: Support Vector Regression-Polynomial; svmRadial: Support Vector Regression -Radial; xgbTree: Extreme Gradient Boosting

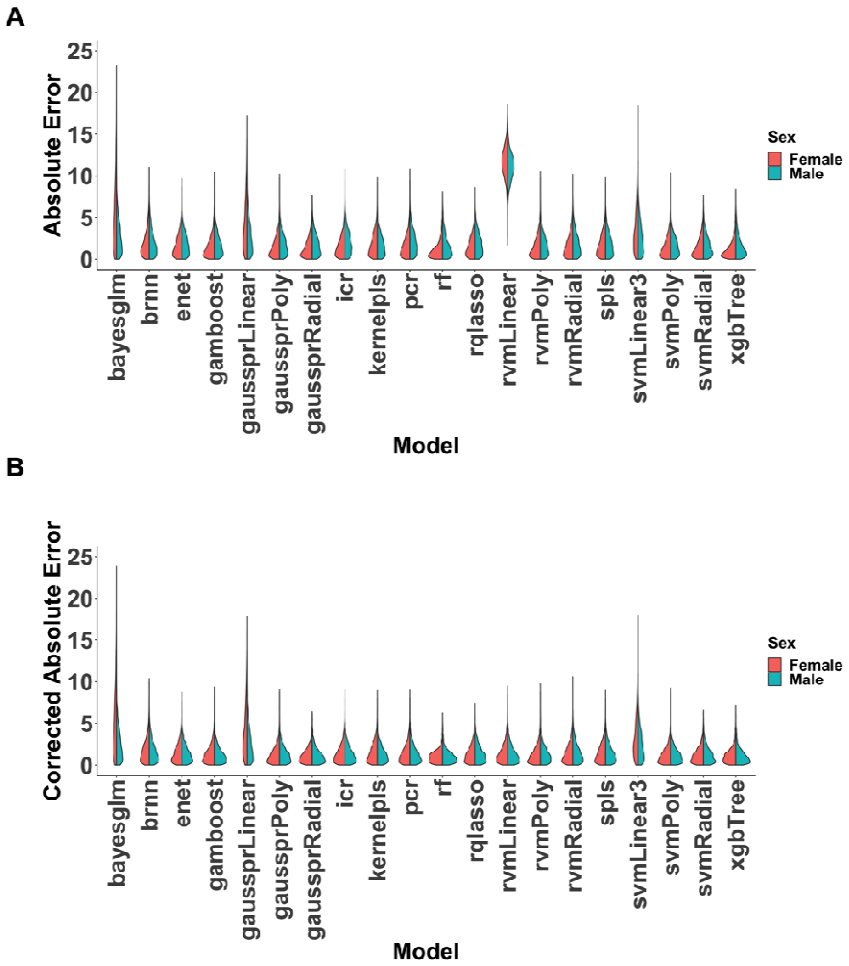

Figure 2. Correlation of predicted age across models. Figure demonstrates correlation between predicted age as estimated by different models in the cross-validated set. **A.** Females; **B.** Males. bayesglm: Bayesian Generalized Linear Model; brnn: Bayesian Regularized Neural Network; enet: Elastic Net Regression; gamboost: Generalized Additive Model with Boosting; gaussprLinear: Gaussian Processes Regression Linear; gaussprPoly: Gaussian Processes Regression Polynomial; gaussprRadial: Gaussian Processes Regression Radial; glm: Generalized Linear Model; icr: Independent Component Regression; kernelpls: Kernel Partial Least Squares; pcr: Principal Component Regression; rf: Random Forest; rqlasso: Quantile Regression with LASSO penalty; rvmLinear: Relevance Vector Machine-Linear; rvmPoly: Relevance Vector Machine-Polynomial; rvmRadial: Relevance Vector Machine-Radial; spls: Sparse Partial Least Squares; svmLinear3: Support Vector Regression-Linear; svmPoly: Support Vector Regression-Polynomial; svmRadial: Support Vector Regression -Radial; xgbTree: Extreme Gradient Boosting

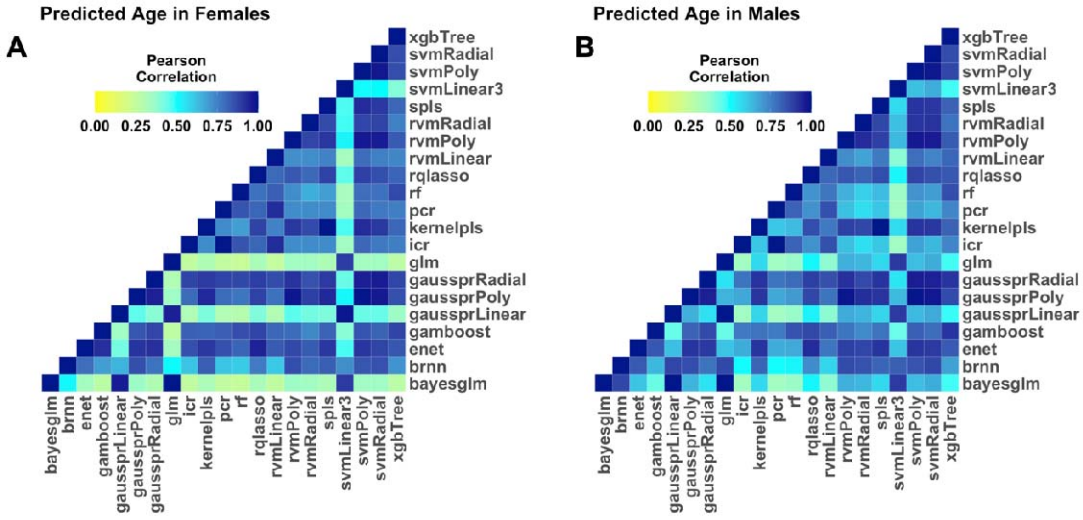

Figure 3. Correlation of chronological age with predicted age, BrainAGE, and bias corrected BrainAGE. Figure shows correlation of chronological age with predicted age, BrainAGE, and bias corrected BrainAGE in the three top-performing models. **A.** Correlation between chronological and predicted age in females; **B.** Correlation between chronological and predicted age in males; **C.** Correlation between chronological age and BrainAGE in females; **D.** Correlation between chronological age and BrainAGE in males; **E.** Correlation between chronological age and bias corrected BrainAGE in females; **F.** Correlation between chronological age and bias corrected BrainAGE in males. rf: Random Forest Regression; svmRadial: Support Vector Regression-Radial; xgbTree: Extreme Gradient Boosting

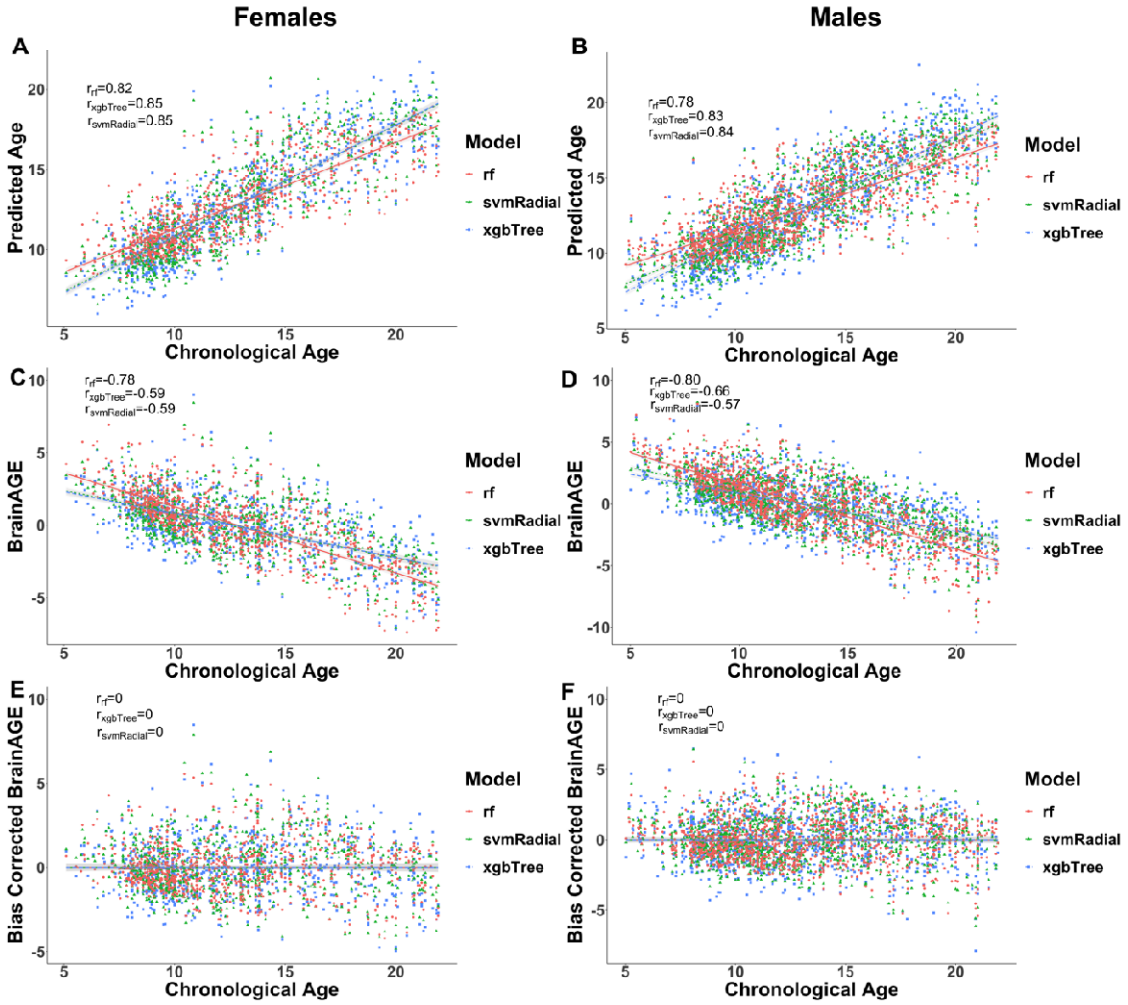

Figure 4. Variable importance. Figure shows variable importance in the three top-performing models superimposed on the 400-parcel Schaefer Atlas. Variable importance values were rescaled such that the variable with the maximum average absolute shapley value in each model has a relative importance of 100. **A.** Variable importance in females; **B.** Variable importance in males. rf: Random Forest Regression; svmRadial: Support Vector Regression-Radial; xgbTree: Extreme Gradient Boosting

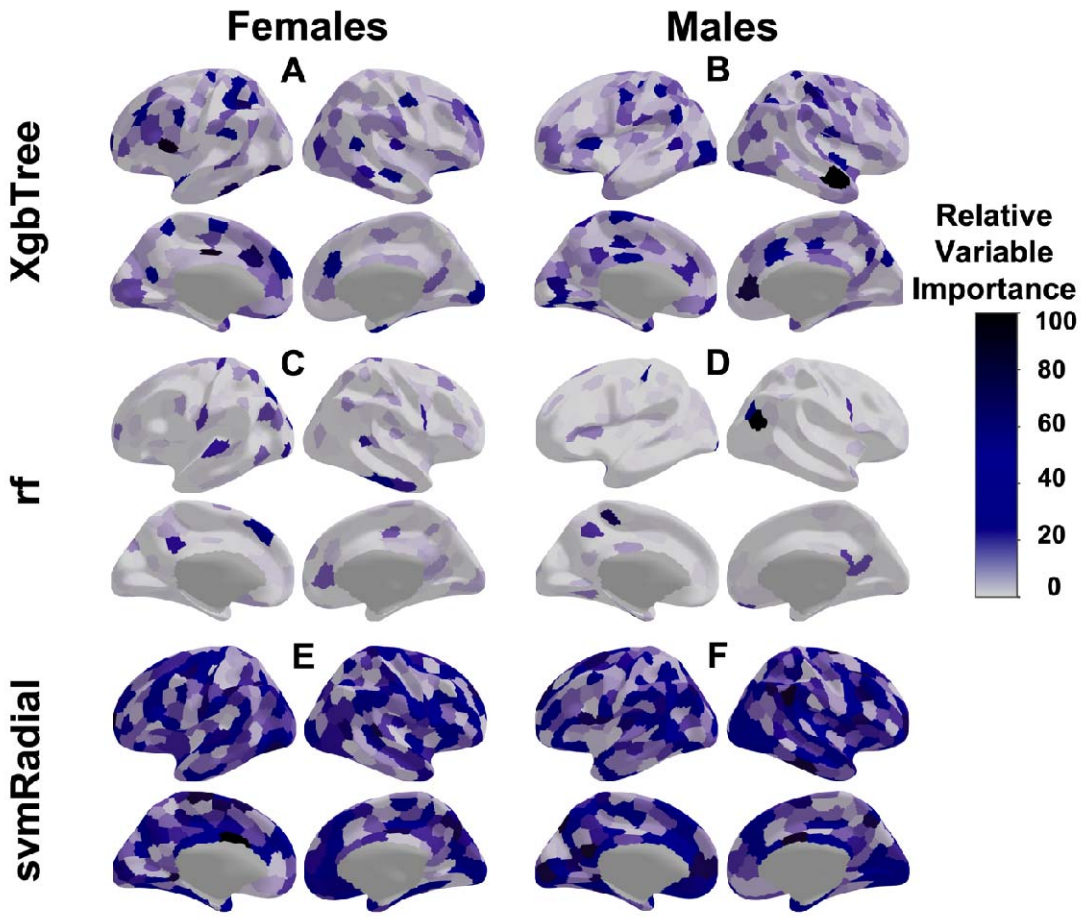
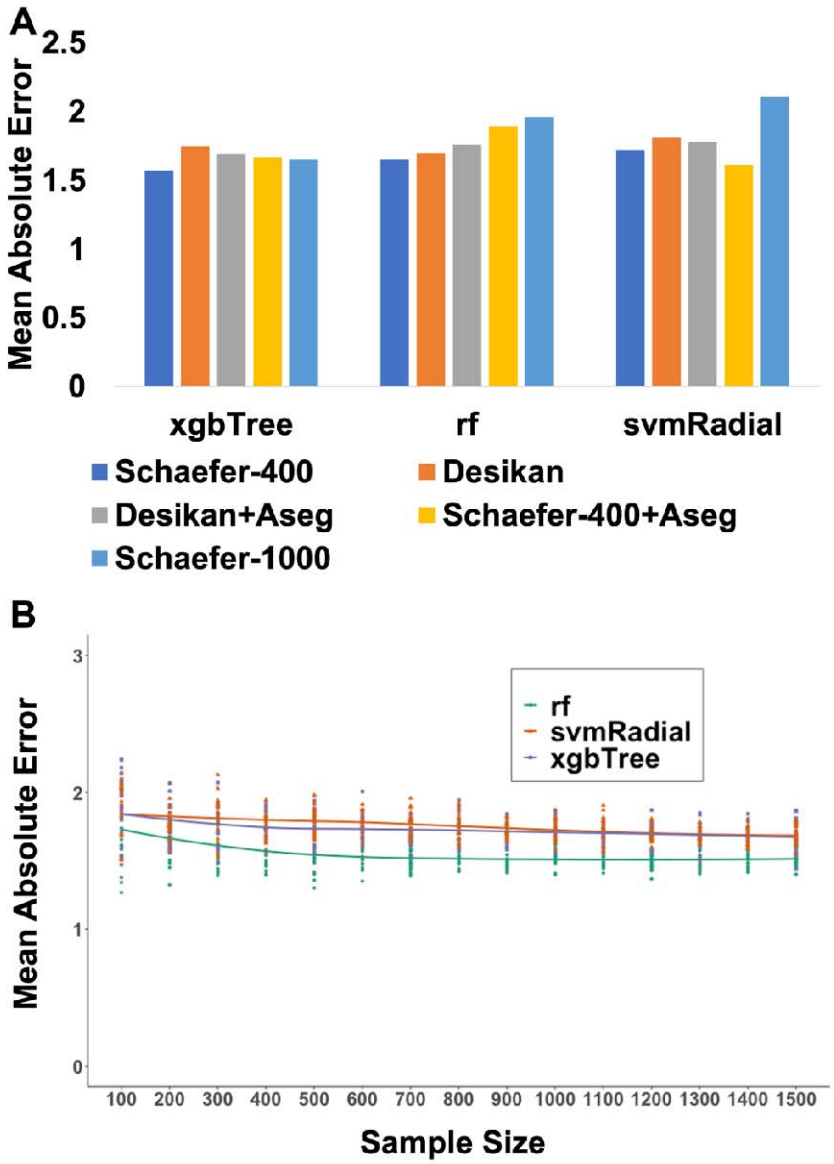
Figure 5. Effect of parcellation scheme and the effect of sample size on model performance. A. Effect of parcellation scheme on model performance of the top three performing models in males (for females see supplemental figure 6); **B.** Effect of sample size on model performance in the three top-performing models The training set was randomly resampled without replacement with a sample size 100 to 1500 (each 20 times); the model was then tested on the hold-out test set. rf: Random Forest Regression; svmRadial: Support Vector Regression-Radial; xgbTree: Extreme Gradient Boosting