Universiteit Leiden

Master Computer Science

Using machine learning to identify brain patterns related to autistic traits

Name: Olga Veth
Student ID: 2067579

Date: 02/08/19

Specialisation: Bioinformatics

1st supervisor: Lu Cao
2nd supervisor: Eduard Klapwijk
3d supervisor: Lara Wierenga

Master Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)

Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Using machine learning to identify brain patterns related to autistic traits

Olga Veth

May 2020

1 Abstract

Individuals with ASD have shown deviation in brain development as well as in brain structure. Multiple studies have used machine learning models in order to predict the age of autistic individuals and compare these with their chronological age. Most deviation occurs within males with ASD, however, it is unknown whether this applies to purely autistic traits instead of ASD diagnosis. Using four regression models, we attempt to predict the so called 'brain age' of a twin dataset (7-11 years old). After optimization using Grid Search and 10-fold cross-validation, three of the four regression model performed with a decent Mean Squared Error. We compared 'brain age error', difference between 'brain age' and chronological age, with their Social Responsiveness Scale scores (measurement of autistic traits). Our results suggest that higher SRS scores are more frequent in males and are predicted to have younger brain age than their chronological age. We found higher positive 'brain age error' values in males in contrast to females. In other words, brain age of males tend to deviate more than their female counterpart if they display mild autistic traits.

2 Introduction

Autism spectrum disorder (ASD) is a common neurodevelopmental disorder and as 'spectrum' implies, it can be diagnosed based on a range of severity of autistic symptoms[1]. In addition to the different levels of severity of autistic traits, ASD also varies in the number of symptoms. The core features of ASD include deficits in social behaviour, such as communication and interaction with other individuals. These symptoms often occur early. As these symptoms suggest, ASD has been associated with atypical brain development[2]. However, due to the heterogeneous nature of ASD, it is difficult to pinpoint specific symptoms since these vary for each individual. Additionally, studies have suggested a complex coherence between genetics, environmental factors and brain physiology in ASD development[2; 3; 4]. In order to shed light upon the complex coherence of factors within ASD, we attempt to relate changes in brain structure during a time period to the heterogeneous behaviour among and between the sexes. In addition, we aim to verify higher variance among males than females taking the heterogeneity into account.

Moreover, age have been found to be significantly correlated with certain signs of ASD, i.e. repetitive behaviour[5]. Similarly, age has been found to be another factor in the heterogeneity in ASD as the symptoms tend to improve over time[6]. Remarkably, a majority of ASD studies have stated a higher prevalence of ASD in males instead of females. One of them estimated the ratio of females with ASD to male to be ranging between 2.7:1 and 7.2:1 in the U.S.A.[7]. Difference in frequency of ASD based on sex indicates that it plays a crucial role in its severity. Whilst no interaction between age and sex have been found[8], both factors have proven to contribute to the heterogeneous nature of ASD. In our study, we are using longitudinal data which enables us to capture the developmental brain structure changes and relate them to the ASD severity. Combination of age as well as sexes provides us to not only find sex-related differences and their relationship with ASD severity but also the structure differences in a given timespan. Although, environmental effects are not taken into account, this complex interaction of age and sex can give us more insight into the heterogeneous behaviour of ASD.

Brain structure have been shown to deviate between non-ASD and ASD individuals. On top of deviation, different brain areas are affected during the lifespan of an ASD individual. Especially, brain development is essential during the childhood of an individual. Studies[8; 9] attempt to understand which structural brain changes occur between individuals with ASD and typically developing children. Structural areas known to be affected are inferior frontal gyrus, the folding of the brain and pars opercularis, a language-related structure. Both areas are reduced in cortical volume and cortical thickness, respectively. Alterations have been found in the brain regions responsible for communication, such as pars opercularis[10] and left rostral middle frontal gyrus[9]. Cortical thickness was found to be lower in the pars opercularis, and the cortical surface area in the inferior frontal gyrus[9]. Moreover, specific communication-related areas in ASD individuals have been found to contain more abnormalities than in typically developed children[11]. Besides the communication-related areas, other brain structures have been suggested as indication of ASD if their structures deviate from the norm. These areas are amygdala, basal, ganglia, frontotemporal lobe, anterior cingulate cortex, and the hippocampus[12]. In short, brain structures of individuals with ASD differs with that of typically developing children and can be detected in scans.

Besides comparison of ASD individuals and typically developing children, sex differences have also been taken into account. As aforementioned, ASD tend to differ between males and females. Studies have shown young females with ASD to have an increased grey brain matter volume in right anterior cingulate cortex and bilateral frontal regions. In contrast, in young males with ASD it was an increased grey matter volume in the left middle occipital gyrus and right superior temporal gyrus[13]. Patterns have been found that indicate the presence of more outliers within males in comparison to females[14]. The deviation between a male ASD individual in comparison with a typical developing individual tend to be higher than their female counterpart. In order to verify this statement, we need to establish if individuals of a given dataset have ASD and in which severity. However, diagnosis of ASD can not be solely be based on brain structure measurements due to the range in brain structure variability in ASD.

Diagnosis of ASD have proven to be a challenge, especially considering the amount of factors to be taken into account. Multiple attempts have been made to detect the presence of ASD in the form of surveys. However, using these surveys, the severity of ASD remains unknown. As a result, downstream analysis can appear quite heterogeneous, as individuals with severe and mild ASD sub-disorders are grouped together. In order to address this issue, we will be using the Social Responsiveness Scale (SRS)[15]. It is a survey which addresses this problem in a dimensional manner as multiple categories are tested. Furthermore, due to SRS its dimensionality, it is possible to determine the severity of these autistic traits in individuals. These SRS scores in combination with structural brain data enables us to get more insight into detection of autistic traits and their severity.

Combination of prior ASD diagnosed individuals and their brain structures have been used in a linear group comparison fashion in recent studies. However, they have been inconclusive to understanding ASD. This may be due to the large heterogeneity in ASD symptoms. Recently, a brain age model was used in order to understand the heterogeneity of ASD[16]. Using an existing model of cortical thickness development in healthy individuals, they addressed the heterogeneity problem. A recent study[14] has used a different approach to asses developmental differences. They estimated the brain age using the volume and thickness of brain areas, and modeled this against the participants chronological age. They utilized an open-source ASD-related dataset, Autism Brain Imaging Data Exchange (ABIDE) II[17], in order to evaluate their model. Using a regression model, they were able to differentiate between healthy and autistic individuals.

Results using their regression model suggested more deviation between typical developed males and ASD-males than in females. In other words, more outliers have been detected based on brain structures in males diagnosed with ASD in contrast to females. Their results were motivated by assuming deviation between ASD and regular individuals instead of taking the accelerated growth in brainsize into account. They used the so-called normative modeling approach in order to individualize the deviation from the mean typically developing brain instead of comparing it to an average ASD individual. Utilizing the model, they found up to almost 10% of their ASD dataset (N=751) to pose significant cortical thickness abnormalities and with the highest deviation (both larger and smaller) to have ASD. These results suggest normative modeling to be an effective method to go beyond the case-control paradigm. In short, the model is able to determine deviation of an individual‘s brain structure to the normative range. Thus, it is suitable

for neurodevelopmental disorder with a heterogeneous range of symptoms such as ASD, as it is individual-based.

However, they did not use longitudinal data and could not follow the development of ASD. Future or past effects on brain areas and cortical thickness are not included. Due to the lack of longitudinal information, brain development of individuals with ASD could not be traced. However, we attempt to use a longitudinal dataset in order to capture the structural brain ages over time. It might provide us with more insight into the sex-differences in variance of brain structure as the given individual grows. Moreover, Bethelehem et al (2019) applied regression models in order to predict the brain age of individuals. In addition, longitudinal data could increase the accuracy of such models as more information is provided about brain structure and its development.

Based on brain structure, ASD individuals can be differentiated from typical individuals. Especially during brain development, ASD individuals have been found to develop in a different manner. Due to this effect, we test whether outliers in brain structures could indicate the presence and severity of ASD and if there is a clear sex-dimorphism between individuals with extremely mild ASD. A younger or older brain indicates a deviation from regular brain development. Using a longitudinal dataset, we move beyond the typical case control paradigm by testing whether ASD symptoms are associated with both upper and lower deviations. By applying a suitable model, we attempt to build an age prediction model in order to determine the so-called 'brain age error'. We proceed to adopt the term 'brain age error' as a reference to the difference between chronological age and brain age using a regression model. With the deviation and the severity of ASD measured using SRS scores of the individuals, we aim to identify brain patterns related to autistic traits and to see whether higher brain age errors are more common in young males than in young females.

3 Material and Methods

3.1 Participants

The present study used data from the Leiden consotrium on individual development study, conducted at the educational-and developmental department of Leiden University. This dataset is comprised of mono-and dizygote twins of a typical developed population. The participants were measured at two timepoints with their ages measured in months. We converted their age into years by dividing by twelve. Overall, participants of both sexes were 7-9 years old during the first timepoint and 9-11 years old in second one. At both timepoints, the percentage of sex was almost evenly distributed. During timepoint 1, the participants consisted of 49% boys and 51% girls and in timepoint 2, 48% and 52% respectively. An overview of the amount of participants can be seen in Table 1.

3.2 MRI acquisition

T1-weighted images of Leiden consotrium on individual development study participants have been acquired using a Philips Ingenia MR 3.0 Tesla scanner at Leiden University Medical Center. Repetition time of 9.8 ms was used in combination with an echo time of 4.6ms. Each scan consists of 140 slices with a voxel size of 1.17 x 1.17 x 1.2 and a field-of-view set to 224 x 177 x 168 mm. Due to our focus on structural brain features in our study, we chose to use T1-weighted images of the dataset.

Table 1: Overview of Leiden consotrium on individual development study. SD stands for 'Standard Deviation'. During timepoint 2, new participants as well as participants from the first timepoint were measured.

	Timepoint 1	**Timepoint 2**
N scans (before Quality Control)	M = 250; F = 264	M = 218; F = 238
Age range	7-9 Years	9-11 Years
Mean (SD) Age in Years	M= 7.852 (0.639);F= 7.960 (0.700)	M= 9.915 (0.655);F = 9.955 (0.717)
Mean (SD) SRS	M = 4.892 (3.616); F = 4.297 (3.485)	M = 4.892 (3.616); F = 4.332 (3.497)
Mean (SD) Length in cm	130.428 (6.634)	142.835 (7.166)
N scans (after Quality Control)	M = 225; F = 246	M = 129; F = 158

3.3 SRS

Each participant from the Leiden consotrium on individual development study have been assessed with SRS[18; 19] in order to evaluate his/her social behaviour. Participants in our current study have been assessed by their parents or guardians with the 'School Age' format used for individuals between 4 and 18 years old.
The questionnaire consists of 65 questions and each section of SRS has a different focus on the social behaviour of the child. These sections include the following aspects: processing of social information, ability in mutual social communication, autistic absent-mindedness, social anxiety and social awareness. These questions are answered by either the parent or guardian of the child.

Questions used within SRS are in the form of statements. A few examples are:

- Is not well coordinated.
- Expressions on his or her face don't match with what he or she is saying.

- Has an unusually narrow range of interests.

Furthermore, these questions are based on the '4-Point Likert Scale' and each answer is worth a certain amount of points. Answers consist of four different options: Not True, Sometimes True, Often True, Almost Always True. 'Not True' is equal to one point, while 'Almost Always True' is worth four points. Every row consists of a question followed by the answers and extra space to fill in the '4-Point Likert Scale' score. After these are filled out, the total of each response value is determined. It is used in order to score the five categories: Restricted Interests and Repetitive Behaviour, Social Motivation, Social Awareness, Social Cognition and Social Communication. The last category, Social Communication and Interaction, is scored using the total score of Social Motivation, Social Awareness, Social Cognition and Social Communication. After the scores of all five categories have been summed up, the scores are converted to so-called T-scores. These scores indicate the level of ASD-like symptoms of the individual. Based on these scores, the individual can be classified as 'Within normal limits'(T-score < 59), 'Mild Range' (T-score > 60 and < 65), 'Moderate Range' (T-score > 65 and < 75) and 'Severe Range'(T-score > 76). The scores among the participants of the current study varied between 0 and 20.

3.4 Demographical data integration

Demographical survey as well as MRI scans were taken at both timepoints for each individual. In order to measure the level of autism among the participants, SRS test was taken as well. Although, two SRS scores are available for each participant, due to the fact that the most recent SRS might be more actual, we chose to use their second SRS score.

As aforementioned, ASD is heterogeneous and acts differently between sexes in symptoms as well as in brain structure. Moreover, studies have indicated that males with ASD tend to deviate more from a typically developing brain than females. Besides our attempt to validate this pattern, the starting brain volume of both sexes differs. As a result, this difference in brain volume could introduce noise during the training process of machine learning models. Finalizing the prior preprocessing of our dataset, we separated our dataset based on sex.

3.5 Preprocessing

3.5.1 Quality Control

An essential step of image processing of MRI scans is quality control. Due to noise introduced by participant, scanner or data preprocessing inaccurate estimations of brain structures are made. Especially, in children or teens wherein they tend to move more during the scanning process. Furthermore, the focus is higher on brain development and interference can lead to inaccurate measurements and conclusions. In order to prevent these issues, quality control techniques are widely applied. Even though each study has its own defined protocol, usually they have a visual inspection in common. However, with an increase in number of participants, visual inspection is a time consuming task, and is sensitive to human error. Thus, in order to reduce misclassification and inconsistency among interpretation of quality of MRI scans, we make use of a quality control tool named, Qoala-T[(20)]. With the use of Qoala-T, we are able to reduce the disadvantages of visual

inspection by reduction of subjective interpretation of MRI scans and misclassification.

Qoala-T is able to perform quality evaluation in partly automated manner, followed by minor visual inspection. It has been specifically designed to be able to perform quality control on structural data that have been preprocessed by FreeSurfer[21]. It uses a supervised learning method to asses the quality of the provided dataset. Quality measurement can be estimated based on a scoring format divided into four different grades: Failed (Score = 0), Doubtful (Score > 0 and < 30),Good (Score > 30 and < 70) and Excellent (Score > 70). It is worth mentioning that these scores are not definite thresholds. It is recommended to manually check 'Good' and 'Doubtful' scans.

Moreover, Qoala-T has two different methods available: model-based and subset-based. In both methods, Qoala-T makes use of a supervised model, named Random Forest (RF)[22]. Supervised models are able to make predictions of previous unseen data after being trained on labeled data. In order to identify and determine the quality of novel data, a 'trainset' is required. Using the model-based approach in Qoala-T, the labeled trainset used is the BrainTime dataset[23; 24; 25; 26], consisting of 784 T1-weighted scans of individuals between 8–25 years during their first measurement and 10-27 years old during their second. BrainTime already has been checked manually and has been assigned quality scores using Qoala-T. In subset-based, the user is able to use its own dataset instead of the BrainTime trainset. We chose to use the model-based method of Qoala-T as we did not have prior quality scores of our dataset available in order to train on the supervised model.

As a result, quality scores of our dataset were given in a table-like manner. Scores below 30 were excluded from further downstream analysis due to their unreliability. The opposite is true for scores above 70. For individuals with quality scores between 30 and 70, a manual evaluation was performed. Manually checking every slice of every single measurement falling into this category would be highly time consuming. In order to reduce the amount of time used for each scan, we made 14 coronal screenshots of each measurement of every 16 slices. Thereafter, we arranged the screenshots into a grid of 7 by 2 into a single image. We evaluated these MRI scans based on segmentation faults due to noise. One of the major problems to check is whether the grey can be differentiated from the grey matter. Movement of the head during the scanning process can lead to noise as 'rings' can be seen throughout the scans. Due to these rings a different brain area can be left out, and non-brain tissue or dura can be included during segmentation. Moreover, there are cases where the temporal pole is not included in the segmentation due to low brightness. All aforementioned factors influence the score, and can decrease the quality of the scan, resulting into 'Failed' or 'Doubtful' categorization.

3.5.2 MRI Segmentation

In order to identify neuroanatomical features of the brain of the participants, scans were processed using FreeSurfer(version 6.0.0)[21]. The software involves a set of different tools necessary in order to perform segmentation on T1-weighted MRI scans. These tools include prior image preprocessing (e.g. removal of noise or non-brain tissue) and measurements of volumes and areas of brain structures. We made use of their cross-sectional as well as longitudinal pipeline[27] in order to process our dataset. For the purpose of initial brain segmentation of each single individual at both timepoints, we utilized the cross-sectional

pipeline as seen in Figure 1. Furthermore, it is part of the initial step to the longitudinal pipeline. Within the cross-sectional pipeline, MRI scans are converted to isotropic images and into Talairach space, a 3D brain coordinate system. Using the Talairach space, each voxel is assigned with a coordinate based on the Desikan-Killiany atlas[28]. Hence anatomical features of the brain are been measured for each individual, i.e. cortical thickness and surface area. Therefore, 185 different brain structures and areas have been measured of left as well as right hemisphere. Few measurements have been excluded as they did not contribute to either surface area or thickness. After the cross-sectional step, only six out of total 514 individuals (1028 MRI scans) runs during the first preprocessing step of the pipeline failed. With remaining scans, we performed our quality control using Qoala-T.

Scans of sufficient quality are used and taken into the second step of the cross-sectional pipeline. These scans are processed within FreeSurfer and matched with the participant itself. Thereafter, we initialized the longitudinal pipeline which shares high similarity with the cross-sectional pipeline but takes into account that these individuals have a certain number of scans and thus, timepoints and a similar brain.

3.5.3 Brain Measurements

In order to grasp the linear curve of the brain development of the participants, we used the advantage of the dataset being longitudinal. We took the mean of both timepoints for the 185 different brain measurements and age of each participant in order to capture the overall brain physiology. In other words, age of each individual is defined as the mean age of that individual of both timepoints. Thereafter, the linear curve is determined by subtraction of the 185 areas of the second timepoint from the first timepoint. This difference provides information in regards to the structural changes in the brain that have occurred within the time span between the two timepoints, i.e. volume of cortical thickness of timepoint 2 - timepoint 1. As age is our predictor variable in our models, we used age to relate to the average measurements as well as the changes in brain structure.

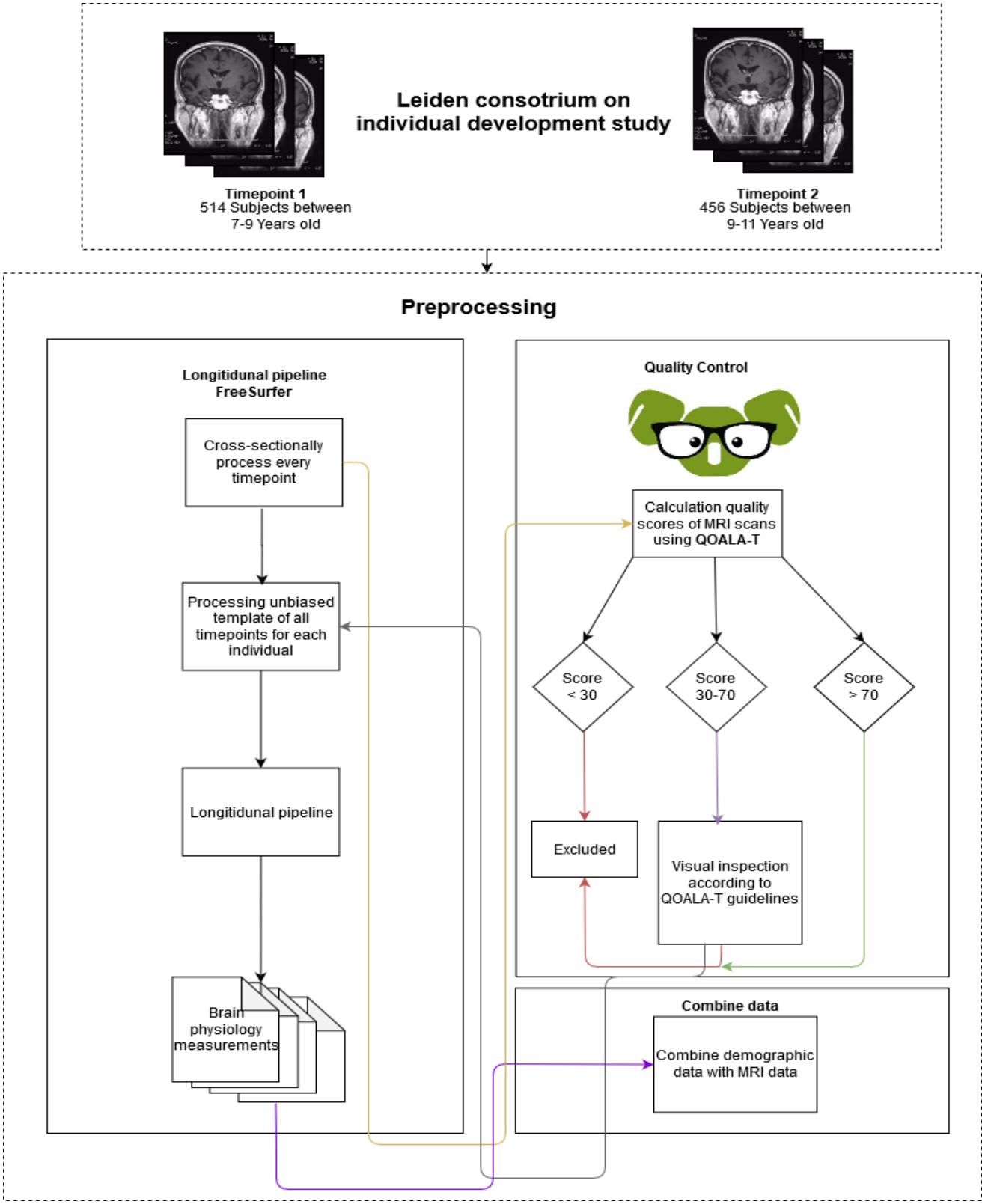

Figure 1: Flowchart showcasing all steps taken during the preprocessing.

3.6 Modeling

3.6.1 Model selection

The aim of our machine learning models is the prediction of age based on brain development and mean measurements of brain structures. As age in the dataset have been measured as values within the range of 7 to 11 years old, our predictive variable is considered to be continuous‘. Model types capable of prediction of continuous variables are so-called regression models. Using regression models, we attempt to perform age prediction on our dataset. However, as multiple regression models exist, we chose the ones that would fit our case.

The first candidate,**Gaussian Process Regression (GPR)**[29] is a non-parametric Bayesian approach. As it is a non-parametric technique, the modeling approach depends on the dataset size for its number of parameters. With its Bayesian properties, it utilizes a probability approach which takes the prior knowledge of event-related conditions for prediction. Last but not least, the GPR is a Gaussian-based approach so it is capable of directly assuming a distribution over functions. Combining the three aforementioned attributes enables a GPR to capture uncertainty of the predicted outcomes as well as perform prediction on continuous variables. Up to 1000 subjects, GPR is able to perform optimal on and as our dataset (N=514) is lower than this threshold, it applies as a model candidate. Moreover, optimizing a GPR can lead to an accurate trade-off of smoothing and fitting the given dataset. However, we have a high number of features (450+) which might decrease GPR's overall performance. Thus, GPR would not be able to tackle its complexity as the dataset is not ‘smooth‘.

A second model we tested, **Neural Network (NN)**[30][31] can extract complicated structure from data due to its ability to learn patterns from a given dataset. These models are based on their biological counterparts as they attempt to ‘learn‘ without a prior set of rules to identify or predict certain outcomes based on their input. Single-layered NN such as the Perceptron[32] are able to solve linearly separable problems. Moreover, multiple kinds of NN exist but as our aim is to predict the age of a given individual using a set of features, we use a deep NN. Deep NN are able to solve non linear and more complex questions. Its self-learning characteristic enables it to accurately make predictions on processed and raw datasets. Moreover, there are no restrictions on the distribution on the dataset. Although, they are capable of ‘learning‘ from the dataset, it comes with the downsides of being computationally intensive and it being a so-called ‘black box‘. Black box indicates the lack of insight of which features have been proven to be the most predictive and correlate with each other during the prediction process of the NN. Moreover, probability of overfitting the dataset is higher using NN and can make it less generalized.

Our third regression model is known for its versatility **Random Forest (RF)**[22]. This model can be used for classification as well as regression. Moreover, it is able to predict with non scaled or even unbalanced datasets. In comparison with the previous two regression models, it uses a straightforward concept, so-called ‘decision trees‘. These decision trees are a tree-like graph, hence their name, wherein each new condition a new ‘branch‘ is made based on whether this criteria is true or false. Probability whether a datapoint falls into a certain category can be determined based on which conditions it passed through. However, these criteria are prone to changes due to the outliers present in the dataset. Combining multiple decision trees makes it more stable against these outliers and more accurate. Therefore, resulting into a Random Forest. Despite that, RF shares

similar problems with NN in its interpretability due to its black box and tend to overfit if its hyperparameters are not optimized.

One type of regression model known to have an accurate performance on high dimensional data is the **Support Vector Regression (SVR)**[33]. SVR has the the goal to minimize the overall error of its predictions. Groups within the dataset are separated with a line surrounded by the widest space possible on one or multiple so-called hyperplanes. A hyperplane is a (sub)space of the original dimension (amount dimensions - 1). If the datapoints are clearly separated on these hyperplanes, the error is minimized. If more data is added into the model, SVR tries to fit them on one of the sides of the given line. However, the lines are sensitive to noise within the data.

Each of these four regression models have their respective disadvantages and advantages. In order to utilize them as optimal as possible we have to optimize their settings, so-called 'hyperparameters'. With the use of task-dedicated subsets of data, train and test, we optimize all four models as will be described in greater detail in the next section. An overview of the complete optimization process can be seen in Figure 2.

3.6.2 Model optimization

We utilized a randomized 70/30 split on the female and male dataset separately to create a train as well as test set. Our train sets consist of 90 subjects and 110 subjects for males and females respectively as seen in Table 2. These values are 70% of their respective dataset size, while their test set is 30%. Therefore, testset contains 39 subjects for males and 48 ones for females. We randomized our train and test set regardless of their family relation due to a lack of standardized method of training using twin data. Each participant was treated as independent individuals. Attributes of the participants, i.e. age and the measurements of the brain areas, have been used in our training and test data.

Table 2: Overview of train-and test set for each sex. Mean as well as number of participants and their mean SRS scores are shown.

	Males		**Females**	
	Train	**Test**	**Train**	**Test**
N scans	90	39	110	48
Mean (SD) age in years	8.794 (0.666)	8.952 (0.611)	8.906 (0.712)	8.891 (0.609)
Mean (SD) SRS	4.821 (3.588)	5.364 (4.699)	4.041 (3.249)	3.769 (2.748)

After preparation of our training and testset, we proceeded to optimization of our four regression models. Due to simplicity and performance differences between packages and coding languages, we chose to make use of R (v.3.5.6)[34] for GPR and RF, and Python (v3.7)[35] for SVR and NN.

In order to increase the performance of our regression models on our dataset as much as possible, we made use of an optimization method. Multiple methods exist for this single purpose but we have chosen Grid Search. As its name implies, it uses a predefined range of values for the variables of the given model, so-called 'hyperparameters' and tries to 'search' for the values with the highest performance. In comparison to the optimization algorithm Random Search[36], it is able to provide better combinations of variables. Although a Grid

Search is able to perform optimization, it is not able to determine the performance of the optimization on its own. It is commonly combined with cross-validation. During cross-validation a subset of the training dataset is randomly splitted up and afterwards used as a separate sub train-and test set for the training process in an iterative manner. We use a 10-fold cross-validation as it gives the most accurate depiction of the performance.
Each of the four chosen regression models are optimized using a 10-fold cross-validation and the same optimization algorithm. However, the steps taken during the optimization are different between models, i.e. scaling of the dataset for SVR. In addition to model-differences, we have practical differences due to packages and distinct programming languages. In order to clarify the variance during optimization process, we will explain the optimization procedure separately for each regression model. However, as GPR and RF share high similarity in the taken steps, they will be discussed in the same section.

3.6.3 GPR & RF

The optimisation procedure for GPR and RF have been done using the package caret (v 6.0-84)[37]. To be more specific, the function 'trainControl' have been used in order to perform the cross-validation as well as Grid Search. In contrast to a regular Grid Search, wherein a predefined range of values is provided of each variable, the function is able to run an automatic grid of these range of values.

Grid Search is present in the same library. Thus, we combined the 10-fold cross-validation without an extra step with the optimization method. Providing Grid Search with out predicted variable, age, and our train dataset, it is able to focus on predicting the brain age. Additionally, the Grid Search function needs to be given a regression model type in order to optimize taken the model into account. Depending on which regression model was used, either 'rf' or 'gaussprPoly' was used, each respectively for RF and GPR with a polynomial kernel. In short, kernels are used in order to map low dimensional data in the form of feature vector of single datapoints against higher dimensional data, so-called 'feature space'. Due to its polynomial characteristic, calculations performed within the kernel are done with addition-like operations.

Following Grid Search, we had to select the hyperparameters of our models. In GPR, we trained two hyperparameters, 'degree' and 'scale'. Changes in 'degree' in GPR determine the linear separability of the kernel. Values of the degree close to '1' are linear and higher values will produce more flexible boundaries between groups within dataset. Meanwhile, 'scale' is used in order to normalize patterns within the dataset, without data modification.
In contrast to GPR, only one variable 'mtry' is optimized in RF. Decision trees splits are based on certain conditions, and they can vary in amount of variables. Hence, we optimized the maximum number of variables that are able to split. Number of trees can also be optimized. However, it is considered to be a performance parameter and less of a tuneable parameter. Accuracy during optimization was measured using Rooted Mean Squared Error (RMSE) but we converted them into Mean Squared Error (MSE) in order to make their performance comparable to the other two models.

3.6.4 SVR

Optimisation of SVR as well as NN have been done in similar manner. Major difference in case of SVR is that it requires an extra step; scaling of the input dataset. In contrast to the commonly used 'StandardScaler', we used a different scaling function to perform feature scaling. Due to our focus on outliers based on brain development, we chose a scaler robust to the effect of outliers. Thereby, 'RobustScaler' was chosen.

Thereafter, the cross-validation was performed using the packages 'scikit-learn'[38]. We used the 'GridSearchCV' function, which as it name implies, performed cross-validation as well as Grid Search optimization. Multiple parameters were chosen to be optimized within GridSearchCV. First one, 'C' determines how much the SVR should avoid misclassification of datapoints by ranging the width of the error range. As 'C' is focused on classification, it is corrected to a more regression-based focus using 'epsilon'. Epsilon is an extra range on top of the error margin, as the width of variance tolerance. Last but not least, the kernel type was optimized based on multiple types, including the aforementioned polynomial kernel.

3.6.5 NN

Although NN does not require an extra transformation step, it has to take other variables into consideration. Beforehand, the NN should be built before optimization can be performed. As it can consist of multiple layers and there has not yet been a guideline for the most effective number of layers for each single problem, we have used one, two and eventually three hidden layers. Using the package Keras (v 2.3.1)[39], we chose an activation function 'relu' for all of our layers, except the output layers which contained a 'linear' activation score. The 'relu' activation function stands for 'Rectified Linear Unit' and is a linear function able to output a positive value or if negative, a zero. Commonly used as a default activation function to a model, it gives the model the advantages of not being computational intensive to train and still have an overall good performance.
The model is compiled with an accuracy measurement of MSE and an 'Adam' optimizer. Due to our dataset comprised more than 400 features, we chose the least computational optimizer, 'Adam' or 'Adaptive moment estimation'. The optimizer shares a high similarity with a different optimizer, stochastic gradient descent. However, while with a stochastic gradient descent, the optimizer is given a single learning rate during optimization, 'Adam' maintains multiple learning rates for each single parameter. Using 'Adam', it is possible to achieve results faster in comparison with a regular stochastic gradient descent.

As the downstream steps require 'Keras', KerasRegressor object is made of the model in order to communicate between 'scikit-learn' and 'Keras'. Optimized variables were 'batch_size' and 'Epochs'. Number of training samples used each time before updating NN model parameters is determined by 'batch_size'. Its second parameter, 'Epochs' determines the frequency of passes through the training dataset.

In 'scikit-learn' we used 'GridSearchCV' function in a similar manner as with the aforementioned SVR. Optimized variables were 'batch_size' and 'Epochs'. Number of training samples used each time before updating NN model parameters are determined by 'batch_size'. Its second parameter, 'Epochs' determines the frequency of passes through the training dataset.

Accuracy of each regression model during optimization was measured with either Rooted Mean Squared Error (RMSE) or Mean Squared Error (MSE). Due to package restrictions, we were only able to determine RMSE in RF and GPR and determined MSE on SVR and NN. Both are accuracy measurements of a regression model, which can be used during validation as well as evaluation. Lower values are considered to indicate accuracy but values close to 0 can indicate overfitting. However, as RMSE is only the rooted version of MSE, we were able to convert the RMSE into MSE. An overview of the Grid Search optimized hyperparameters for the four regression models and their MSE are shown in Table 3.

Table 3: Overview of optimized settings based on Grid Search and 10-fold cross-validation of the four regression model candidates given SamenUniek data

Sex	**GPR**	**NN**	**RF**	**SVR**
Male	Degree: 1 Scale: 0.01	Batch_size: 10 Epochs: 150	Mtry: 66	C: 1.5 Epsilon: 0.1 Gamma: 1e-07 Kernel: linear
	MSE: 0.295	MSE: 1.692	MSE: 0.363	MSE: 0.258
Female	Degree: 1 Scale: 0.01	Batch_size: 5 Epochs: 50	Mtry: 70	C: 1.5 Epsilon: 0.3 Gamma: 1e-07 Kernel: linear
	MSE: 0.377	MSE: 1.107	MSE: 0.449	MSE: 0.290

3.7 Brain age error

We proceed to the calculation of the 'brain age error' (brain age vs. chronological age) after the models predicted the 'age' of the individuals of our dataset. The brain age error is calculated by subtracting the brain age of the chronological age of the individual. In order to relate the deviation in age to behavioural traits of autism, we compared the performance of the participants on SRS to the brain age error. Moreover, it could also give more insight into sex-differences in variance of brain age error. Package ggplot2(v 3.2.1)[40] and plotly (v 4.9.1)[41] were used in order to visualize the brainage prediction and comparison of deviation to SRS scores.

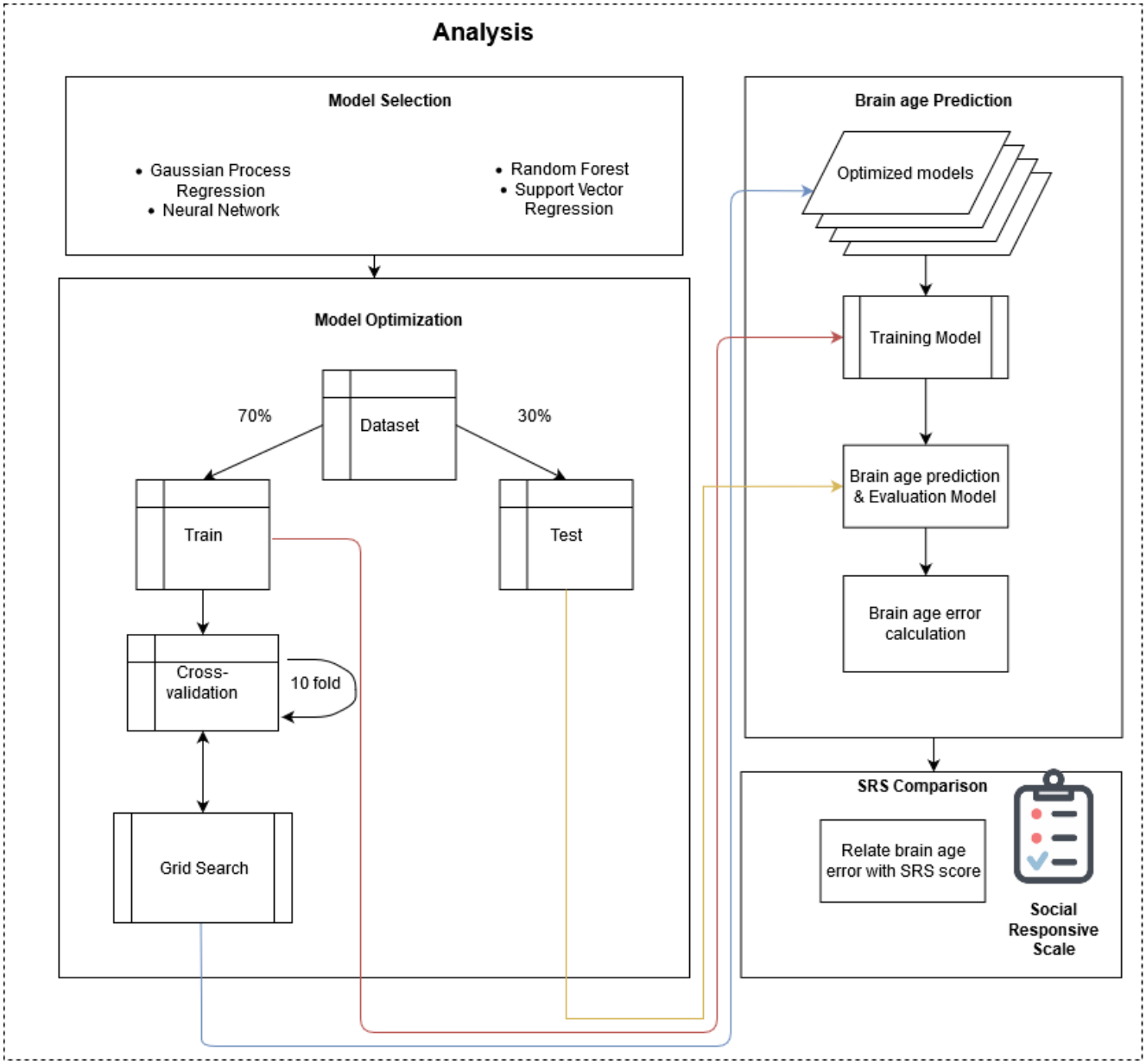

Figure 2: Flowchart showcasing all steps taken during modeling.

4 Results

4.1 Brain age prediction

Four types of regression models (GPR, RF, SVR and NN) have been used in order to predict the brain age of the participants from the Leiden consotrium on individual development study. In order to compare the performance of our four regression models, we used an intercept only linear model as the base line. We predicted their brain age based on physiological brain measurements differences between two timepoints as well as the average. Using the differences, we were able to capture the developmental slope. In order to determine the slope, we only used participants which had two timepoints. Participants measured at only one timepoint have been excluded. Whilst with average measurements we can capture the overall brain features within the time period. Due to aforementioned sex-differences in brain structures such as increased gray matter in males[13; 14] and our focus on sex-based differences, we have separated the dataset based on sex. Thereafter, we determined the brain age error (chronological age - brain age) of each participant.

4.1.1 Brain age prediction of young males

We predicted the brain age of our test dataset (N=39) in males after training the four aforementioned regression models on our training data (N=90). As seen in Table 4, we see that GPR, RF as well as SVR perform better than intercept-only model as their mean brain age ranges between 8.799 and 9.088 years, close to the real mean chronological age of 8.952 years. In contrast to the intercept-only model, which deviates most of the models with a mean brain age of 6.78. Furthermore, intercept-only model shows the highest standard deviation among the models and insignificant negative correlation with age. Additionally, intercept-only model predicts the participants to be at least 1 year older than their chronological age and has much bigger range even though our age range is at most 4 years, as seen in Figure 3. Negative or low correlation between predicted age and chronological age could indicate its predictive underperformance. In addition to its underperformance, intercept-only model has the highest MSE as well as MAE among males as well as female cohorts, as seen in Table 5. As a result, its mean brain age error is at least 2 years with the biggest standard deviation among all models in both cohorts. Low performance of our baseline model could have been due to overfitting the training set. As the number of features is higher than our number of participants, it overfits during training and leads to an underperformance during testing. Even though our baseline model underperformed, our remaining models except NN have shown to perform better based on mean brain age as well as deviation.

Besides the underperformance from intercept-only model, RF and GPR present comparable mean brain age as well as correlation with chronological age, as seen in Table 4. Standard deviation of both regression models are 0.189 and 0.204 and translates into Figure 3 as a lack of range. Both models have in common that they show limitations in age range for their participants, especially in comparison to SVR with a higher standard deviation of 0.294. Brain age predicted for the male participants in both RF and GPR is within the range of between 8.3 and 9.2 years old as seen in Figure 3. Small range in predicted values translates into a horizontal distribution of datapoints, as seen in Figure 3. As a result, individuals are mainly horizontally distributed. Furthermore, their brain age

error as well as accuracy measured using MSE and MAE barely differs. Even though GPR and RF share similar values, they differ in their correlation with SRS scores. Variance in predicted age by GPR versus chronological age is more positively correlated (0.147) with SRS scores of the male participants compared to RF (0.123).

Apart from the similar outcomes of GPR and RF, NN shows in comparison an unique result. Each run of our trained NN gave different results so we were not able to perform accurate measurements of each unique run. We interpreted the performance of NN based on Figure 3. In contrast to GPR, RF and SVR, NN was not able to accurately predict the age of participants as not a single datapoint crosses the X=Y line. All of its predictions have been made underneath our X=Y line. Participants have been clustered together between 7.5 and 8.9 based on brain age, although few of them have a chronological age of 8.9 years old and older. These results imply that NN predicted a younger age for each participant than their chronological age. However, as all participants are predicted to be younger than their chronological age, the accuracy is the lowest among the four regression models with a value of 1.69209 as can be seen in Table 4.

In contrast to the other models, SVR has a negative mean brain age. On average, SVR tend to predict male participants to be older than their chronological age in contrast to the other models. In addition, it has a higher standard deviation compared to GPR and RF which implies a bigger range of predicted brain age, see Table 4. We see the high variance in brain age of male participants between the age of 8.2 years old and 9 years old in Figure 3. Most of the variance exists of positive predicted brain age values. Older predictions are made of these individuals between the given age range and tend to deviate more than the younger brain age counterpart. Nevertheless, a more linear trend is seen in Figure 3 in SVR which is likely due to its performance. Its accuracy can be seen in Figure 3 as multiple datapoints are nearby the red linear line. More linearity indicates a more accurate prediction performance. Additionally, SVR's mean brain age (9.088) is the closest one to the real mean chronological age of the male participants (8.951). Its significant positive correlation with age is lowest among our regression models but close to RF and GPR. Its mean brain age error as well as MAE are the smallest among the models on the male cohort, see Table 4. Based on these two factors, we can conclude that SVR has the highest performance among all regression models on our male cohort.

Table 4: Mean brain age in years as well as *pearson correlation* (**r**) and its *p-value* using the regression models and intercept-only model on participants of the Leiden consotrium on individual development study. Models were tested on the testdata of both genders. **GPR** = *Gaussian Process Regression*, **RF** = *Random Forest*, **NN** = *Neural Network* and **SVR** = *Support Vector Regression*.

		Mean (SD) brain age (years)	**r**	**p-value**
Males (N=39)	**Intercept-only**	6.781 (9.100)	-0.127	0.443
	GPR	8.799 (0.204)	0.625	2.130e-05
	NN	-	-	-
	RF	8.800 (0.189)	0.613	3.355e-05
	SVR	9.088 (0.294)	0.595	6.550e-05
Females (N=48)	**Intercept-only**	8.777 (5.283)	-0.161	0.273
	GPR	8.891 (0.180)	0.474	6.650e-04
	NN	-	-	-
	RF	8.911 (0.167)	0.462	9.558e-04
	SVR	8.868 (0.291)	0.450	0.001

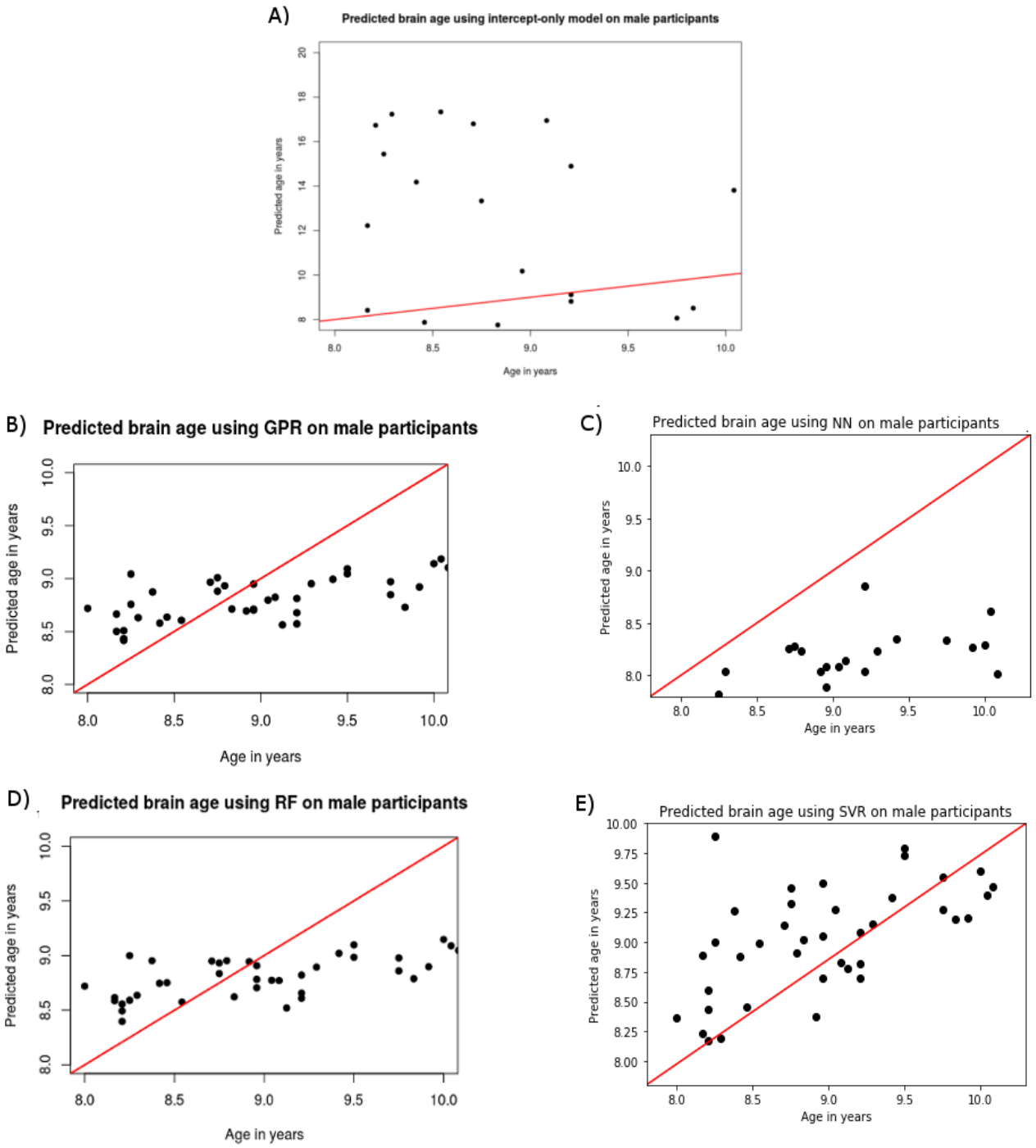

Figure 3: Prediction of age using the four regression models on the testdata of male participants of the Leiden consotrium on individual development study. Models were tested on the testdata of male participants (N=39). (**A**) *Intercept-only model*, (**B**) *Gaussian Process Regression*, (**C**) *Random Forest*, (**D**) *Neural Network* and (**E**) *Support Vector Regression*. Chronological age is shown on the x-axis and brain age on y-axis. Red linear line (X=Y) represents the ideal age prediction to compare with the four models.

4.1.2 Brain age prediction of young females

Brain age prediction using our regression models was also performed on the females after training on our train set (N=110) and testing on the test set (N=48). In contrast to males, intercept-only model has a mean predicted age close to the real mean age as seen in Table 4. However, the remaining three regression models still perform better with a mean brain age of between 8.868 and 8.911 versus the chronological mean age of 8.891. The intercept-only model has a lower standard deviation than its counterpart in males but still over-and underestimates the age of female participants as our dataset's maximum years span is 4 years. Especially participants between 8 and 9 years old are overpredicted in age, see Figure 4. Moreover, its predicted brain age is similarly as in males slightly negatively correlated with chronological age. Its MSE in female is lower compared to the males cohort but is still higher than our cohort's age range of 4 years. This does not apply to its MAE which is slightly lower than the age range. Intercept-only model's overall mean brain age error as well as its standard deviation is lower (0.113) compared to the male datast (2.171), see Table 5. Due to aforementioned reasons, these results imply that its low performance on the cohort lead to low insignificant correlation. Our three remaining regression models do perform better in comparison to the intercept-only model.

Comparable to the male cohort, RF and GPR share similarities in their overall performance on the female cohort.Firstly, both models present similar mean brain age with a minor difference of 0.020 years. Although, the difference is higher than in males, both models tend to have a similar age prediction. Secondly, their standard deviation implies their predicted age range to be between 0.167 and 0.180, according to Table 4. Their correlation with chronological age are both positive, insignificant and only differ 0.012. Bigger contrast is their significance, as GPR is more significant in brain age correlation than RF. Last but not least, their MSE differs only with 0.010 years, as seen in Table 4. Their MSE as well as MAE are visualized as an overall horizontal distribution of the datapoints in Figure 4. These results indicate an overestimation of age of younger individuals and underestimation of older individuals. However, RF has a negative brain age error which suggests that it tend to predict female participants to be older than their chronological age, see Table 5. In contrast to GPR which has a positive brain age error which is close to 0. Even though GPR has a brain age error close to 0 and lowest MSE, it does not have the lowest MAE. However, based on its mean brain age error, significance correlation and MSE, GPR has the best results on the female cohort among our models.

In contrast to the NN's performance with the male participants, datapoints are shown both left and right of the X=Y line. Moreover, its datapoints deviate between 7.8 and 9.5 years old. In comparison to RF and GPR, NN is able to predict a bigger range of values. However, this does not indicate a higher accuracy. Based on its MSE of 1.10745, it has as in the male cohort, the worst performance of the four regression models. Participants of 10 years of age are predicted to be between 7.9 years old or 9 years old. Individuals younger than 9.5 years old tend to be more accurately predicted as they follow the linear line.

While SVR was the most accurate model in male cohort, it slightly underperforms GPR on the female cohort. Its mean brain age deviates the most from the real mean chronological age besides the intercept-only model in Table 4. However, its standard deviation does indicate its ability to predict a bigger range of values as potential brain age. Based on Figure 4, we see that SVR is able to predict participants younger than the age of

9.5 years old more accurately, as they follow a more linear trend. Moreover, correlation between its predicted brain age with chronological age is close to RF and both present the same p-value. Still, SVR performs slightly better based than RF on its MSE which is 0.302. Whilst SVR has a better MAE, GPR does prove itself to be most accurate based on MSE and mean brain age error.

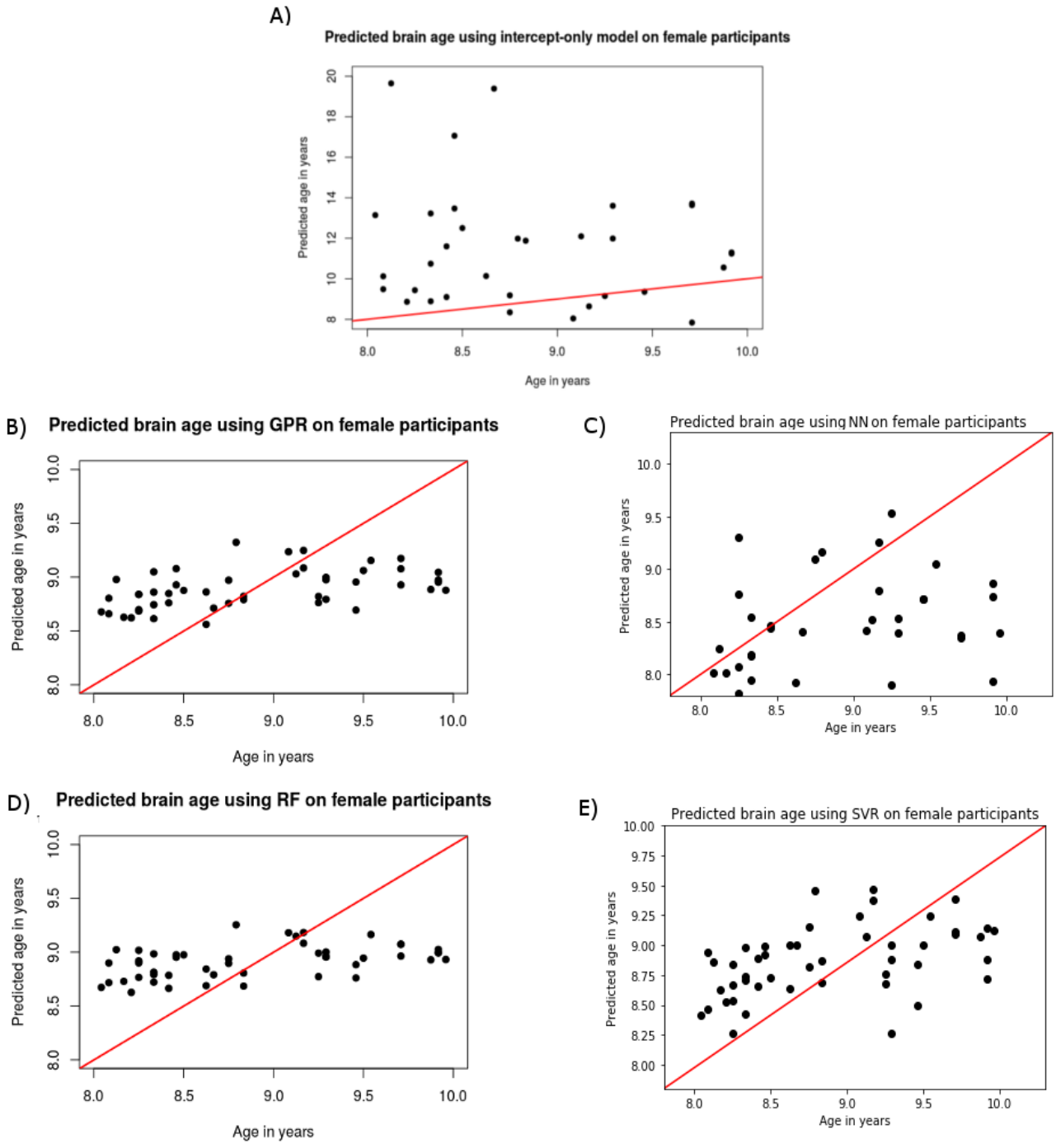

Figure 4: Prediction of age using the four regression models on the testdata of female participants of the Leiden consotrium on individual development study. Models were tested on the testdata of female participants (N=39). (**A**) *Intercept-only model*, (**B**) *Gaussian Process Regression*, (**C**) *Random Forest*, (**D**) *Neural Network* and (**E**) *Support Vector Regression*. Chronological age is shown on the x-axis and brain age on y-axis. Red linear line (X=Y) represents the ideal age prediction to compare with the four models.

Table 5: Overview of brain age error and its correlation with SRS scores of participants from the Leiden consotrium on individual development study. Accuracy was measured using (**MSE**); *Mean Squared Error* and (**MAE**); Mean Absolute Error. The following four regression models have been used: **GPR** = *Gaussian Process Regression*, **RF** = *Random Forest*, **NN** = *Neural Network* and **SVR** = *Support Vector Regression*. In order to determine correlation and its significance, we used *Pearson correlation* **r** and **p-value**

		MSE (years)	MAE (years)	Mean (SD) brain age error (years)	r	p-value
Males (N = 39)	**Intercept-only**	87.115	7.446	2.171 (9.196)	0.192	0.243
	GPR	0.276	0.441	0.153 (0.509)	0.147	0.415
	NN	1.692	-	-	-	-
	RF	0.283	0.448	0.152 (0.517)	0.123	0.496
	SVR	0.258	0.421	-0.075 (0.509)	0.189	0.292
Females (N = 48)	**Intercept-only**	28.725	3.926	0.113 (5.415)	0.269	0.273
	GPR	0.292	0.467	-3.890e-04 (0.547)	-0.184	0.262
	NN	1.107	-	-	-	-
	RF	0.302	0.477	-0.024 (0.555)	-0.191	0.245
	SVR	0.290	0.459	0.023 (0.544)	-0.100	0.546

4.2 Brain age error in relation to SRS

Our aim is to see whether variance in brain age error is more prevalent in males or in females with autistic traits. We used the SRS scores of the participants and related them with the brain age error of our regression models. Whether these brain age errors are related to SRS scores is determined by correlation and its significance.

4.2.1 Relation between SRS and brain age error in young females

Our base-line models had the biggest mean brain age error and as aforementioned, highest MSE. Moreover, it does present similar to the other regression models, variance in brain age error in participants with a SRS score below 10, see Figure 5. In contrast to the other models, the highest SRS scoring male participants are predicted to be overall older than their chronological age as their brain age error is negative. Taken these into account, intercept-only model does present highest correlation with SRS. Although, it is insignificant. In the case of GPR as well as RF, we see five participants with an SRS score higher than 12 in Figure 5. The brain age error of the male participants is between 0 and 1 years, so both models predicted male participants to be younger than their chronological age as can be seen in their positive values in Table 4. Lower scoring participants with an SRS score below 10, have either a positive or negative brain age error. Although, negative brain age error shows less variation as datapoints tend to cluster more than their positive counterpart. Due to lack of variance in negative brain age errors, it leads to a more positive average brain age error. Beside the overall trend, one participant with an SRS score of 0 (complete lack of autistic-traits) shows to have almost 0 brain age error in GPR as well as in RF. Even though the brain age errors of GPR and RF are slightly positive correlated with SRS scores, they are not significant as their p-values are above 0.05.

No negative brain age errors are present on the y-axis of NN due to the previous predictions of solely younger predicted brain age in comparison to their chronological age. Participants with a SRS score below 8 appear to fall between 0.2 and 2 as brain age error.

Only one participant with a SRS score of between 1 and 2 has a higher brain age error of 3. Above the SRS score of 10, one participant seems to have a high brain age error of 2.5. It implies a younger brain age that could be related to a higher SRS score.
In the case of SVR, the range of positive brain age error (0.7) is smaller than the negative brain age error (-1.7). This is represented in its mean brain age error which is negative in contrast to the other models which have a positive brain age error, according to Table 5. All except one participant with an SRS score of below 10 appear to show more positive than negative brain age error. The single outlier with a SRS score of 2 shows the most negative brain age error. A negative brain age error indicates an older predicted brain age compared to its chronological age. Participants with a SRS score higher than 10 have a positive brain age error of bigger than 0. Although, SVR's brain age errors are slightly more correlated than GPR and RF, the correlation is insignificant.

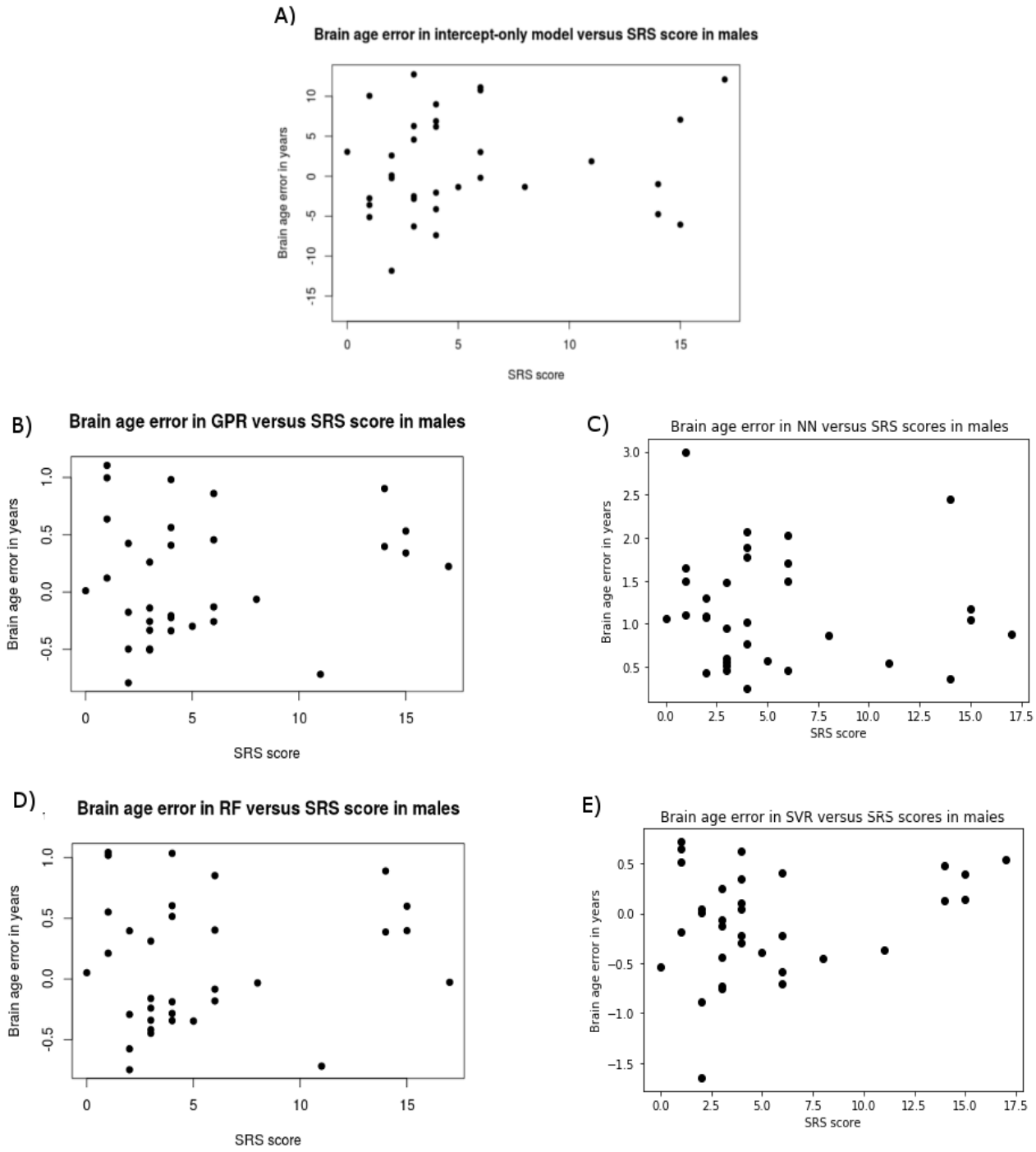

Figure 5: Brain age error of the four regression models versus SRS scores of male participants of the Leiden consotrium on individual development study. (**A**) *Intercept-only model*, (**B**) *Gaussian Process Regression*, (**C**) *Random Forest*, (**D**) *Neural Network* and (**E**) *Support Vector Regression*. SRS scores of participants is shown on the x-axis and brain age error on y-axis.

4.2.2 Relation between SRS and brain age error in young females

Similarly as in males, intercept-only model varies in its age prediction among participants scoring lower than 9 on the SRS test. However, in contrast to the other regression models,

the higher scoring participants have negative brain age errors. These results imply an older predicted age than their chronological age. In addition, intercept-only model has a lower mean brain age error as well as standard deviation, see Table 5. Its correlation with SRS is higher than in the males cohort but still not considered to be significant. Continuing the similarities between GPR and RF, their brain age error and relating them to SRS scores are similar as can be seen back in Figure 6. For example, individuals with SRS scores below 8 present positive as well as negative brain age error scores. Two participants with a SRS score above 10 show positive values of brain age error of between 0 and 0.3. Similar pattern can be seen in SVR as well as NN with a sole difference of higher brain age errors in NN. These two participants might have a younger brain age compared to their chronological age. Both GPR and RF their brain age errors are negatively correlated with SRS scores. Their correlation differs only with 0.015. Single significant correlation is found in GPR with a p-value of 0.034 which is below our limit of 0.05.
In contrast to the male participants, the brain age error scores in NN occur negative as well as positive. Majority of the participants below SRS score of 8 have a positive brain age error, except one participant with an SRS score of 1 with a brain age error of -1.2. Based on these values, these individuals appear to have a younger brain age.
Similarly to RF and GPR, most individuals in SVR with a SRS score below 8 are distributed between positive and negative brain age error and higher scoring individuals have a brain age error between 0 and 0.5. Each model shares an almost similar pattern of brain age errors predicted for higher SRS scoring female participants. Furthermore, SVR has the lowest negative insignificant correlation among the regression models, see Table 5.

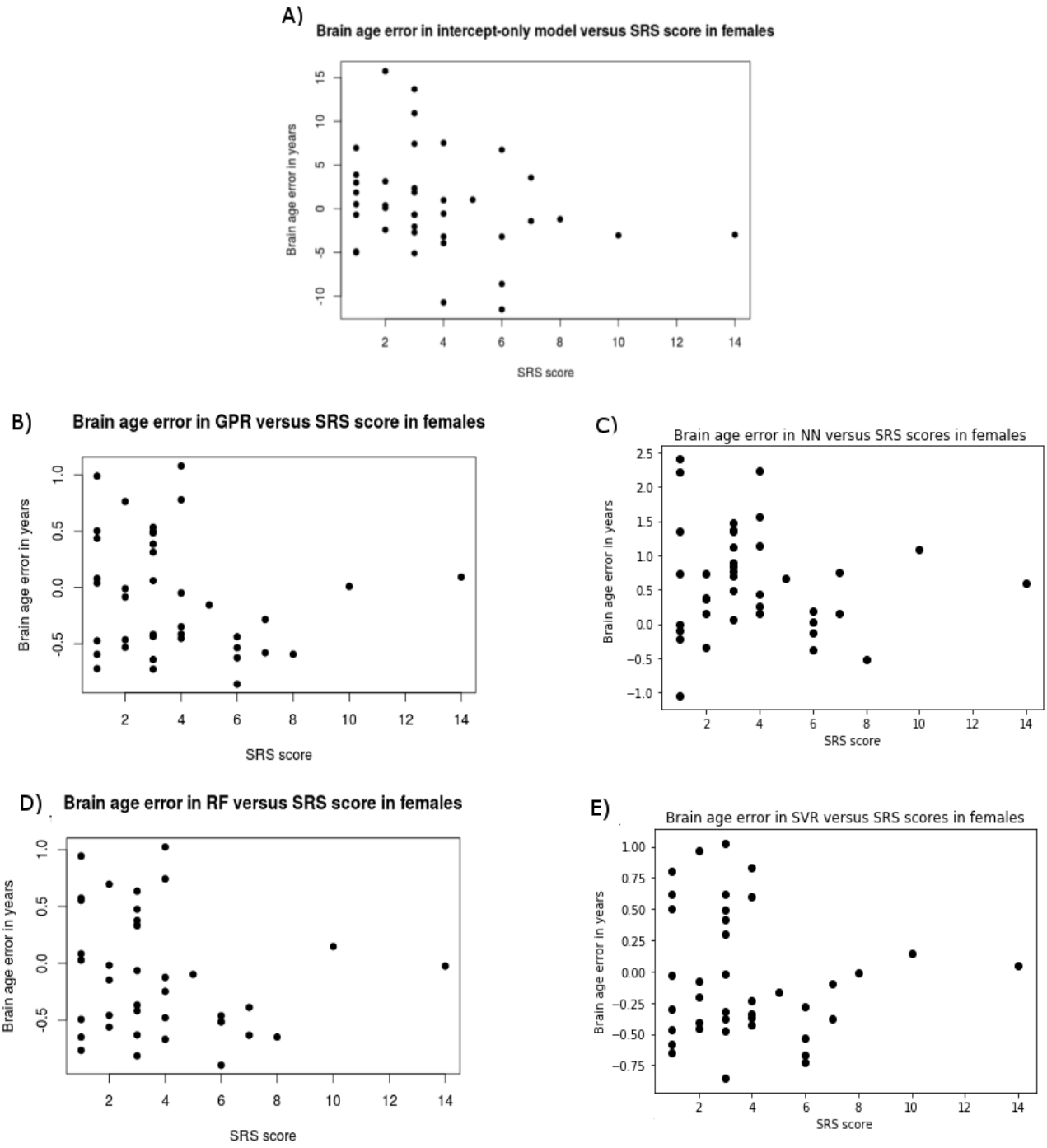

Figure 6: Brain age error of the four regression models versus SRS scores of female participants of the Leiden consotrium on individual development study. (**A**) *Intercept-only model*, (**B**) *Gaussian Process Regression*, (**C**) *Random Forest*, (**D**) *Neural Network* and (**E**) *Support Vector Regression*. SRS scores of participants is shown on the x-axis and brain age error on y-axis.

5 Discussion

Our objective was to observe brain patterns that are related to autistic traits and test the hypothesis that outliers in overall brain development are more prevalent in males compared to females. In order to detect these brain patterns, we predicted the brain age with regression models using structural brain data from a twin dataset. Using multiple accuracy measerements, SVR was the most accurate on the male cohort and GPR on the female cohort. Males were predicted to be on average younger than their real age. In order to relate the severity of autistic traits in the form of SRS scores, with their brain development, we calculated their brain age error. We did not find significant correlation between brain age error and the SRS scores. In addition, our results did imply more variance among males as their mean brain age error was higher than their female counterparts.

Different optimal models are found within our results for both sexes, GPR for females and SVR for males. Their MSE was the lowest among the models and their mean brain age were the closest to the mean chronological age of each of the respective sexes. Our sex-differentiated models tie well with previous studies[42; 43] wherein they trained sex-specific brain age model and applied them to the same as well as opposite sex. In Erus et al (2015), T1 images of typically developed participants (8-22 years old) were segmented with an in-house pipeline. With SVR they predicted a MAE of 1.291 years on males and female-model with MAE of 1.264 years on females. Their accuracy decreased when these sex-models were used on the opposite sex. The male-models overestimates brain age of females and vice versa, especially among the younger individuals within their dataset. If sex differences in brain models were neglectable, sex-specific models would have predicted the same values in both sexes. In short, studies[42; 44] have suggested that sex difference is prominent in brain development and might have lead us to different optimal models for each sex.

Additionally, we found specific model types to be more accurate in certain sexes. Most studies [45; 46; 47; 43] wherein they corrected for the sex-related differences reported that GPR outperformed SVR. In Niu et al (2019), GPR performed better in both sexes compared to SVR. However, this difference was smaller in males than in females with a difference of 0.010 years in MAE on a dataset within the age range of 8-21 years[43]. In contrast to our study, they did not use MSE or an extra accuracy measurement. Even though MAE was in both sex cases the lowest in GPR, it can not be concluded whether GPR did overall perform better than their SVR. However, difference in performance based on MAE have been minor. This is in line with our results which show small differences in performance between GPR and SVR for both sexes. This slight difference can be due to multiple factors, including the used dataset. Whether the specific model type is related to the slightly better performance in a certain sex has yet to be explored.

Based on our results, variance in males are higher compared to females but no model showed positive significant correlation with SRS scores. Lack of correlation between brain age error and severity of autism among majority of our models is in line with other studies [48; 49]. In Tunç et al (2019), they found no significant differences in correlation between ASD severity and brain age error using sex-models. With a young age-range of 6-25 years old, they trained an SVR on Freesurfer segmented T1 scans in order to predict brain age as well as our so-called 'brain age error' on typically developed participants. A similar

conclusion was reached by Hus et al (2012) wherein they yielded no sex-differences on the raw SRS scores of same-sex siblings between 6 and 11 years old. Nonetheless, one of the SVR models in Tunç et al (2019) was correlated with brain age error. Participants with older predicted brain age showed the least ASD severity and younger predicted brain age more severity. These findings are directly in line with our high-scoring participants (SRS score > 9) which are the most so-called 'severe' ASD cases within our dataset and were predicted, independed of sex, to have younger predicted brain age. In short, age plays a more important role than sex in correlation of predicted age with severity of autistic traits.

Limitations of our study include the lack of utilization of our twin dataset. During this study, we treated our twins as family-independent individuals and neither correct or used the similarities and environmental factors. Our focus was not on the brain development differences between twins but rather between 'regular individuals'. Both types of twins share high similarity in their genetic background and translation into their brain physiology[50; 51; 52]. In order to correct for the impact of environmental as well as genetic factors between twins, Structural Equation Modeling (SEM)[53] can be used. SEM uses a combination of regression principles and factor analysis and have been applied in twin studies[47] during training process of models. Implementation of SEM could provide us with insight into the sex-differences and variance in brain development as well as relationship with autistic traits of both twin types.

As aforementioned, most brain age studies [47; 54] have been done on participants with adults or older cohorts. Age ranges used were at least 25 years, higher than our maximum of 4 years. Regression models used in these models have performed with low MAE on the given cohort. They have been able to capture the aging process as well as maturation process of adult and elderly individuals. As these processes are captured at different phases, more distinct patterns may be found with their models. In addition, more distinct pattern leads to a higher accuracy in age prediction. However, our participants' overall brain maturation process is measured in a smaller time lapse. While changes in brain structure occur among children within our age range, these are less distinguishable for regression models than in the aforementioned studies. Although, it is uncertain to conclude whether and how much the performance of our regression models have been affected by lack of age range. In short, increasing number of timepoints or age range might improve overall accuracy of our models.

In regards with improvement of our models, we utilized Grid Search as our optimization algorithm. Multiple optimization methods exist. Two optimization techniques for SVR introduced in Cherkassky et al (2004)[55] and Chalimourda et al (2004)[56] have been used in Franke et al (2010)[57] instead of Grid Search and 10-fold cross-validation. They motivated their choice of optimization technique as it would be less computationally expensive and in order to produce sample-dependent parameters. Both techniques resulted with an accurate MAE of between 4 and 5 years in a dataset consisting of two datasets of individuals between 19-86 years old and 20-59 years old. Our SVR can be optimized using the discussed algorithms to see whether its performance improves on both cohorts.

Whilst we used an optimization algorithm, our regression models still performed with overall low accuracy. One of the most common potential explanations is the used number of features. We did not exploit feature selection and reduction in our study. Removal of features during feature selection can lead to a better understanding of their overall

influence and contribution on the regression model. Furthermore, partially relevant features can provide noise, introduce overfitting and more computational expensive process of training and testing a model. One example is Sparse Group Lasso (SGL)[58; 59]. It is able to detect groups within features that are related to the outcome variable and select the most influential features within those groups. Aycheh et al (2018)[60] have used SGL to improve the MAE of their GPR from 4.063 to 4.053. However, considering the used age range is between 45 and 91, the improvement is minor. Despite the fact that these techniques reduce noise and increase accuracy, it can decrease the performance of models. Certain regression models have been designed to handle high dimensional dataset and use that information. Reducing the number of features could impair such models significantly. Although there is a probability of reducing the performance of our regression models, it is worth to expand on.

Our RF and GPR models predicted the brain age of younger individuals to be older and vice versa. This phenomenon is named 'systemic bias' and has reoccurred in multiple age prediction models[47; 61; 62]. Systemic bias tend to cause noise within brain age prediction and its error calculation. Associating the brain age error with cognitive or psychological disorders can be highly influenced due to bias. Recent study[46] has shown that systemic bias is unrelated to imbalance of sample size as well as potential noise within the dataset. In order to test whether systemic bias is purely due to data or type of prediction model, Liang et al (2019), used GPR, NN, SVR and Ridge Regression on healthy participants (N=2026) between 6-89 years old. All four models presented systemic bias within their age prediction on their FreeSurfer segmented scans. As of now, cause of this phenomenon is still unknown but studies have bypassed it by utilizing linear mixed models[43; 63]. Linear mixed model have been used in order to correct for the bias as well as sex differences[63]. In future work, linear mixed models can be used in order to correct for the systemic bias in age prediction of our models.

Increasing in popularity as a brain age prediction framework is brainAGE[57]. It utilizes a machine learning method named Relevance Vector Machine[64] in order to make an age estimation based on structural MRI data of given individual. Advantage of brainAGE is its ability to capture the brain maturation of children as well as adults and elderly. BrainAGE has proven to be scanner-independent and does not decrease in its age prediction performance if different scanners have been used. As of now, brainAGE performed with a decent accuracy among different studies with varying goals[65; 66; 67]. However, we did not use brainAGE during our study as our focus lies on testing models and while brainAGE is based on a regression model, it is considered to be a framework. Even though we did not apply brainAGE in our study, it might be worth exploring whether its prediction accuracy would be on our twin dataset and if similar patterns with SRS would be found.

In conclusion, we were able to create age prediction models for a young twin dataset. Based on accuracy, SVR had the best performance on males and GPR in females. Using these models, we found variance to be more prevalent in males compared to females according to average brain age error and deviation. We found positive brain age errors in individuals with high SRS scores (>9) in both sexes but these were not significantly correlated. Based on younger predicted age as well as higher variance in males, our results indicate that

bigger brain development deviation might be related to mild autistic traits.

References

[1] D. H Geschwind and P. Levitt, "Geschwind DH, Levitt P. Autism spectrum disorders: developmental disconnection syndromes. Curr Opin Neurobiol 17: 103-111," *Current opinion in neurobiology*, vol. 17, pp. 103–111, 2007.

[2] E. Courchesne, "Brain development in autism: Early overgrowth followed by premature arrest of growth," *Mental retardation and developmental disabilities research reviews*, vol. 10, pp. 106–111, 2004.

[3] K. Wang, H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. A. Sleiman, C. E. Kim, C. Hou, E. Frackelton, R. Chiavacci, N. Takahashi, T. Sakurai, E. Rappaport, C. M. Lajonchere, J. Munson, A. Estes, O. Korvatska, J. Piven, L. I. Sonnenblick, A. I. Alvarez Retuerto, E. I. Herman, H. Dong, T. Hutman, M. Sigman, S. Ozonoff, A. Klin, T. Owley, J. A. Sweeney, C. W. Brune, R. M. Cantor, R. Bernier, J. R. Gilbert, M. L. Cuccaro, W. M. McMahon, J. Miller, M. W. State, T. H. Wassink, H. Coon, S. E. Levy, R. T. Schultz, J. I. Nurnberger, J. L. Haines, J. S. Sutcliffe, E. H. Cook, N. J. Minshew, J. D. Buxbaum, G. Dawson, S. F. A. Grant, D. H. Geschwind, M. A. Pericak-Vance, G. D. Schellenberg, and H. Hakonarson, "Common genetic variants on 5p14.1 associate with autism spectrum disorders," *Nature*, vol. 459, p. 528, apr 2009.

[4] S. N. Basu, R. Kollu, S. Banerjee-basu, and B. Lane, "AutDB : a gene reference resource for autism research," vol. 37, no. November 2008, pp. 832–836, 2009.

[5] A. J. Esbensen, M. M. Seltzer, K. S. L. Lam, and J. W. Bodfish, "Age-Related Differences in Restricted Repetitive Behaviors in Autism Spectrum Disorders," *Journal of Autism and Developmental Disorders*, vol. 39, no. 1, pp. 57–66, 2009.

[6] P. T. Shattuck, M. M. Seltzer, J. S. Greenberg, G. I. Orsmond, D. Bolt, S. Kring, J. Lounds, and C. Lord, "Change in Autism Symptoms and Maladaptive Behaviors in Adolescents and Adults with an Autism Spectrum Disorder," *Journal of Autism and Developmental Disorders*, vol. 37, no. 9, pp. 1735–1747, 2007.

[7] M. W. Report, "Prevalence of Autism Spectrum Disorders — Autism and Developmental Disabilities Monitoring Network , 14 Sites , United States , 2008," vol. 61, no. 3, 2012.

[8] M. A. Hattier, J. L. Matson, K. Tureck, and M. Horovitz, "The effects of gender and age on repetitive and/or restricted behaviors and interests in adults with autism spectrum disorders and intellectual disability," *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2346–2351, 2011.

[9] L. E. Libero, T. P. DeRamus, H. D. Deshpande, and R. K. Kana, "Surface-based morphometry of the cortical architecture of autism spectrum disorders: volume, thickness, area, and gyrification," *Neuropsychologia*, vol. 62, pp. 1–10, 2014.

[10] S. Yamasaki, H. Yamasue, O. Abe, M. Suga, H. Yamada, H. Inoue, H. Kuwabara, Y. Kawakubo, N. Yahata, S. Aoki, Y. Kano, and N. Kato, "Reduced Gray Matter Volume of Pars Opercularis Is Associated with Impaired Social Communication in

High-Functioning Autism Spectrum Disorders," *BPS*, vol. 68, no. 12, pp. 1141–1147, 2010.

[11] E. Redcay, "The superior temporal sulcus performs a common function for social and speech perception: Implications for the emergence of autism," *Neuroscience & Biobehavioral Reviews*, vol. 32, no. 1, pp. 123–142, 2008.

[12] D. G. Amaral, C. M. Schumann, and C. W. Nordahl, "Neuroanatomy of autism," *Trends in Neurosciences*, vol. 31, pp. 137–145, mar 2008.

[13] A. Retico, A. Giuliano, R. Tancredi, A. Cosenza, F. Apicella, A. Narzisi, L. Biagi, M. Tosetti, F. Muratori, and S. Calderoni, "The effect of gender on the neuroanatomy of children with autism spectrum disorders: a support vector machine case-control study," *Molecular Autism*, vol. 7, no. 1, p. 5, 2016.

[14] R. A. I. Bethlehem, J. Seidlitz, R. Romero-Garcia, G. Dumas, and M. V. Lombardo, "Normative age modelling of cortical thickness in autistic males," *bioRxiv*, p. 252593, jan 2019.

[15] J. N. Constantino, S. A. Davis, R. D. Todd, M. K. Schindler, M. M. Gross, S. L. Brophy, L. M. Metzger, C. S. Shoushtari, R. Splinter, and W. Reich, "Validation of a Brief Quantitative Measure of Autistic Traits : Comparison of the Social Responsiveness Scale with the Autism Diagnostic Interview-Revised," vol. 33, no. 4, 2003.

[16] M. Zabihi, M. Oldehinkel, T. Wolfers, V. Frouin, D. Goyard, E. Loth, T. Charman, J. Tillmann, T. Banaschewski, G. Dumas, R. Holt, S. Baron-cohen, S. Durston, S. Bölte, D. Murphy, C. Ecker, J. K. Buitelaar, C. F. Beckmann, and A. F. Marquand, "Archival Report Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 4, no. 6, pp. 567–578, 2019.

[17] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, pp. 659–667, jun 2014.

[18] F. J. Aldridge, V. M. Gibbs, K. Schmidhofer, and M. Williams, "Investigating the Clinical Usefulness of the Social Responsiveness Scale (SRS) in a Tertiary Level, Autism Spectrum Disorder Specific Assessment Clinic," *Journal of Autism and Developmental Disorders*, vol. 42, no. 2, pp. 294–300, 2012.

[19] J. N. Constantino, "Social Responsiveness Scale BT - Encyclopedia of Autism Spectrum Disorders," pp. 2919–2929, 2013.

[20] E. T. Klapwijk, F. V. D. Kamp, M. V. D. Meulen, S. Peters, and L. M. Wierenga, "NeuroImage Qoala-T : A supervised-learning tool for quality control of FreeSurfer segmented MRI data," *NeuroImage*, vol. 189, no. January, pp. 116–129, 2019.

[21] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, 1999.

[22] T. K. Ho, "Random Decision Forests," p. 278, 1995.

[23] E. Schreuders, B. R. Braams, N. E. Blankenstein, J. S. Peper, B. Güroğlu, and E. A. Crone, "Contributions of Reward Sensitivity to Ventral Striatum Activity Across Adolescence and Early Adulthood," vol. 89, no. 3, pp. 797–810, 2018.

[24] E. A. Crohne, "Unraveling age, puberty and testosterone effects on subcortical brain development across adolescence," vol. 91, pp. 105–114, 2018.

[25] S. Peters and E. A. Crone, "Increased striatal activity in adolescence benefits learning," vol. 8, no. 1, p. 427, 2017.

[26] A. I. Becht, M. G. N. Bos, S. A. Nelemans, S. Peters, W. A. M. Vollebergh, S. J. T. Branje, W. H. J. Meeus, and E. A. Crone, "Goal-Directed Correlates and Neurobiological Underpinnings of Adolescent Identity: A Multimethod Multisample Longitudinal Approach," 2018.

[27] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis," no. 2011, 2012.

[28] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.

[29] M. N. Gibbs, "Bayesian Gaussian Processes for Regression and Classi cation,"

[30] S. Kleene, "Representation of Events in Nerve Nets and Finite Automata," vol. 34, 1956.

[31] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, pp. 2554 LP – 2558, apr 1982.

[32] F. Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Report: Cornell Aeronautical Laboratory, Cornell Aeronautical Laboratory, 1957.

[33] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, V. Vapoik, W. Long, and B. Nj, "Support Vector Regression Machines w) tw ,," vol. 1.

[34] R. Team, "A language and environment for statistical computing," *Computing*, vol. 1, 2006.

[35] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[36] N. Baba, "Convergence of a random optimization method for constrained optimization problems," *Journal of Optimization Theory and Applications*, vol. 33, no. 4, pp. 451–461, 1981.

[37] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software, Articles*, vol. 28, no. 5, pp. 1–26, 2008.

[38] F. Pedregosa, R. Weiss, and M. Brucher, "Scikit-learn : Machine Learning in Python," vol. 12, pp. 2825–2830, 2011.

[39] F. Chollet *et al.*, "Keras." `https://github.com/fchollet/keras`, 2015.

[40] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[41] C. Sievert, *plotly for R*, 2018.

[42] 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 (New York, N.Y. : 1991)*, vol. 25, pp. 1676–1684, jun 2015.

[43] X. Niu, F. Zhang, J. Kounios, and H. Liang, "Improved prediction of brain age using multimodal neuroimaging data," *Human Brain Mapping*, vol. 41, pp. 1626–1643, apr 2020.

[44] M. S. Goyal, T. M. Blazey, Y. Su, L. E. Couture, T. J. Durbin, R. J. Bateman, T. L.-S. Benzinger, J. C. Morris, M. E. Raichle, and A. G. Vlassenko, "Persistent metabolic youth in the aging female brain," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 3251 LP – 3255, feb 2019.

[45] A. Xifra-porxas, A. Ghosh, G. D. Mitsis, and M.-h. Boudrias, "Estimating brain age from structural MRI and MEG data : Insights from dimensionality reduction techniques 1 Introduction," pp. 1–29, 2019.

[46] H. Liang, "Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders," no. November 2018, pp. 3143–3152, 2019.

[47] J. H. Cole, R. P. K. Poudel, D. Tsagkrasoulis, M. W. A. 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.

[48] B. Tunç, L. D. Yankowitz, D. Parker, J. A. Alappatt, J. Pandey, R. T. Schultz, and R. Verma, "Deviation from normative brain development is associated with symptom severity in autism spectrum disorder," pp. 1–14, 2019.

[49] V. Hus, S. Bishop, K. Gotham, M. Huerta, and C. Lord, "Factors influencing scores on the social responsiveness scale," *Journal of Child Psychology and Psychiatry*, vol. 54, pp. 216–224, feb 2013.

[50] D. Posthuma and D. I. Boomsma, "A Note on the Statistical Power in Extended Twin Designs," *Behavior Genetics*, vol. 30, no. 2, pp. 147–158, 2000.

[51] T. White, N. C. Andreasen, and P. Nopoulos, "Brain Volumes and Surface Morphology in Monozygotic Twins," *Cerebral Cortex*, vol. 12, pp. 486–493, may 2002.

[52] W. F. C. Baaré, H. E. Hulshoff Pol, D. I. Boomsma, D. Posthuma, E. J. C. de Geus, H. G. Schnack, N. E. M. van Haren, C. J. van Oel, and R. S. Kahn, "Quantitative Genetic Modeling of Variation in Human Brain Morphology," *Cerebral Cortex*, vol. 11, pp. 816–824, sep 2001.

[53] L. J. Eaves, K. A. Last, P. A. Young, and N. G. Martin, "Model-fitting approaches to the analysis of human behaviour," *Heredity*, vol. 41, no. 3, pp. 249–320, 1978.

[54] B. Mwangi, K. Hasan, and J. Soares, "Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach," *NeuroImage*, vol. 75, jul 2013.

[55] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, no. 1, pp. 113–126, 2004.

[56] A. Chalimourda, B. Schölkopf, and A. J. Smola, "Experimentally optimal ν in support vector regression for different noise models and parameter settings," *Neural Networks*, vol. 17, no. 1, pp. 127–141, 2004.

[57] K. Franke, G. Ziegler, S. Klöppel, and C. Gaser, "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters," *NeuroImage*, vol. 50, no. 3, pp. 883–892, 2010.

[58] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, vol. 22, pp. 231–245, apr 2013.

[59] S. Chatterjee, "Sparse Group Lasso : Consistency and Climate Applications,"

[60] H. M. Aycheh, J.-K. Seong, J.-H. Shin, D. L. Na, B. Kang, S. W. Seo, and K.-A. Sohn, "Biological Brain Age Prediction Using Cortical Thickness Data: A Large Scale Cohort Study," *Frontiers in Aging Neuroscience*, vol. 10, p. 252, 2018.

[61] H. M. Aycheh, J.-K. Seong, J.-H. Shin, D. L. Na, B. Kang, S. W. Seo, and K.-A. Sohn, "Biological Brain Age Prediction Using Cortical Thickness Data: A Large Scale Cohort Study," *Frontiers in Aging Neuroscience*, vol. 10, p. 252, 2018.

[62] H. R. Pardoe and R. Kuzniecky, "NAPR: a Cloud-Based Framework for Neuroanatomical Age Prediction," *Neuroinformatics*, vol. 16, no. 1, pp. 43–49, 2018.

[63] Y. Chung, J. Addington, C. E. Bearden, K. Cadenhead, B. Cornblatt, D. H. Mathalon, T. McGlashan, D. Perkins, L. J. Seidman, M. Tsuang, E. Walker, S. W. Woods, S. McEwen, T. G. M. van Erp, T. D. Cannon, and N. for the North American Prodrome Longitudinal Study (NAPLS) Consortium and the Pediatric Imaging and Genetics (PING) Study Consortium, "Use of Machine Learning to Determine Deviance in Neuroanatomical Maturity Associated With Future Psychosis in Youths at Clinically High Risk," *JAMA Psychiatry*, vol. 75, pp. 960–968, sep 2018.

[64] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, sep 2001.

[65] 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.

[66] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, and A. D. N. Initiative, "BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease," *PloS one*, vol. 8, pp. e67346–e67346, jun 2013.

[67] K. Franke, C. Gaser, and for Initiative, "Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease," *GeroPsych - The Journal of Gerontopsychology and Geriatric Psychiatry*, vol. 25, pp. 235–245, dec 2012.