Can your brain tell us how old you are?

Machine learning answers to a fundamental question in neuroscience

J. Asher, K.T. Dang, P. Klopfenstein, M. Masters, and J. Yeater

Supervisor: Prof. A.M. Selvitella

Abstract

Our study aimed to find a new method for predicting the age of a subject from their brain biometrics (brain age problem). We utilized an isoperimetric-type ratio T^2/SA, where T is the membrane thickness and SA is the surface area, for each region of the brain. This was done to create an explainable model and as a dimensionality reduction technique. We trained a multivariable regression model on the ratios, but it failed to preserve the prediction accuracy of the multivariable regression model trained on the area and thickness data. Therefore, we tried other modeling techniques, utilizing a multinomial logistic model. When compared to the multinomial logistic model trained on the thickness and area data, the multinomial logistic model trained on the ratios preserved, and improved, prediction accuracy. Therefore, the isoperimetric-type ratios can be used as an explainable dimensionality reduction technique for modeling age using brain biometrics.

1. Introduction

Predicting age of a patient based on quantitative measurements of the brain is of much interest to modern medicine. Among these quantitative measurements including the volume, surface area, and thickness of specific gray matter regions as measured by an MRI. To predict age based on these measurements, we consider the fact that in adults, brain regions tend to shrink with age, and the corrugation of the brain tends to change as well. These changes are also affected by physical or mental health conditions; however, this paper will not take them into consideration.

The Human Connectome Project (HCP) is a government funded initiative to map the functions of the human brain. The goal of the project is to uncover patterns to explain the functions of the brain related to human's behaviors. In our research, our concentration is the relationship between the volume, surface area, and thickness of different regions and the age. The HCP is a large data repository. To support our research, we will use only the "HCP Young Adult" dataset. This set contains data from patients aged 22-35, and 14 individuals aged 36 and above. It will be a challenge to predict age based on young adult data, because there are many fluctuations during this early stage. The data is also restricted; therefore, we do not have the exact ages of the patients, only the groups of ages to which the patients belong. Most studies on this topic will have the specific ages of the patients at their disposal.

1.1 Background literature

Physical measurements of the brain as measured by an MRI are of great interest to medical researchers for their potential in predicting aging and illness. Specifically, changes in gray matter are known to manifest in normal aging, dementia, ALS, Huntington's disease, and schizophrenia. One of the measurements we are interested in is the thickness of the gray matter of various brain regions, which can vary greatly depending on the region (Fischl 2000). It is important to note that there are many different methods used to measure the thickness of the gray matter. These methods include closest point methods, which finds the minimal distance between a point on the inner surface of the gray matter to a given point on the outer surface, as well as Laplace methods, which solve Laplace's equation for the potential between the inner and outer surface, which gives a more complicated measure. There is also a method that finds the radius of the largest inscribed sphere in the gray matter (Thorstensen 2014).

Previous studies have successfully used machine learning to predict the age of a patient given brain measurements. Using the proper algorithms, Al Zoubi et. al. have been able to predict age with the lowest mean absolute error of 6.87 years using only EEG signals. Their model used 500 subjects, and extracted specific features from the EEG signals which were then given to the machine learning algorithm. They also used k-fold cross validation to create a general linear model. A study from Seoul proposed a model using cortical thickness that was able to predict age with a mean absolute error of 4.05 years. Non-robust statistical measures were used, so outliers were filtered out. Techniques used include sparse group lasso, gaussian process regression, deep neural networks, and cross validation. Their most successful model was a hybrid between the sparse group lasso and the gaussian process regression (Aycheh).

1.2 Dataset description

We are using the 1200 subject release behavioral dataset from the Human Connectome Project. There are 1206 subjects in total, with 247 in the 22-25 age range, 527 in the 26-30 age range, 418 in the 31-35 age range, and 14 in the 36+ age range. The results are gathered from different imaging modalities, in addition to behavioral and genetic data. More specifically, according to HCP S1200 Release Reference Manual, 1113 subjects have 3T MR structural scans, 889 subjects fulfil all the four 3T MRI modalities, which are structural images (T1w and T2w), resting-state fMRI (rfMRI), task fMRI (tfMRI), and high angular resolution diffusion imaging (dMRI). Additionally, 3T HCP protocol Retest data are also provided with 46 subjects and these subjects are all monozygotic twins, 21 twin pairs and 4 MZ twins without retest of co-twin. In addition to 3T MR scan, 184 subjects also have 7T MR scan data, along with 95 subjects which also have at least some resting-state MEG (rMEG) and/or task MEG (tMEG) data, available.

This study mainly concentrates on the category of FreeSurfer of the dataset. FreeSurfer is a brain imaging software package. It contains programs whose main focus is to analyze MRI scans of brain tissue. Its tools are useful for conducting any volume-related and surface-related analysis because, with the ability to reconstruct models of both the gray/white and pial surfaces, it can help measure cortical thickness, surface area and folding.

In this research, the data of the volume, the area and the thickness of different regions of the brain obtained from FreeSurfer category will be taken into consideration when making the prediction of the age of the subjects.

The regions of the brain which were used for calculating the models are listed in the appendix.

1.3 Isoperimetric ratio

Suppose $\Omega \subset \mathbb{R}^n$ is a connected region with smooth boundary $\partial \Omega$. We define $V(\Omega)$ to be the volume of Ω and $S(\Omega)$ to be the surface of Ω. Given this two quantities, we can compute the isoperimetric quotient as

$$IQ(\Omega) = \frac{S(\Omega)^n}{V(\Omega)^{n-1}}.$$

This quantity can be defined for every dimension n. For example, for $n=3$, we recover the 3D volume and its surface, while for $n=2$, we recover the 2D area and perimeter.

The minimum of IQ is reached when Ω is the unit ball B_1 in $\mathbb{R}^n$ and we have the so called isoperimetric inequality:

$$IQ(\Omega) \geq n* V(B_1).$$

In particular, for $n=2$ we have:

$$IQ(\Omega) \geq 4\pi.$$

In the HCP dataset used in this paper, we have information about surface area and thickness; however, we do not have complete information about the perimeter of the region, only of the thickness $T(\Omega)$. We will therefore use a surrogate for the isoperimetric ratio, given by

$\tilde{IQ}(\Omega)=\frac{T(\Omega)^2}{S(\Omega)}.$

1.4 Research question

Our study aims to understand whether a subject's age can be predicted from their brain biometrics. Specifically, whether the isoperimetric-type ratio (ITR) defined by T^2/A can be used as a dimensionality reduction technique to create an explainable model.

2. Methods

2.1 Preliminary analysis

In order to determine what variable reduction techniques to use, we plotted the correlation coefficients between different variables and age, in the decreasing order. This was done with all of the freesurfer data, as well as the dimensionality reduction with thicknesses and areas combined to form the ITRs. Since no graphs showed any clear cutoff in the correlations, we decided to use all of the variables in our models.

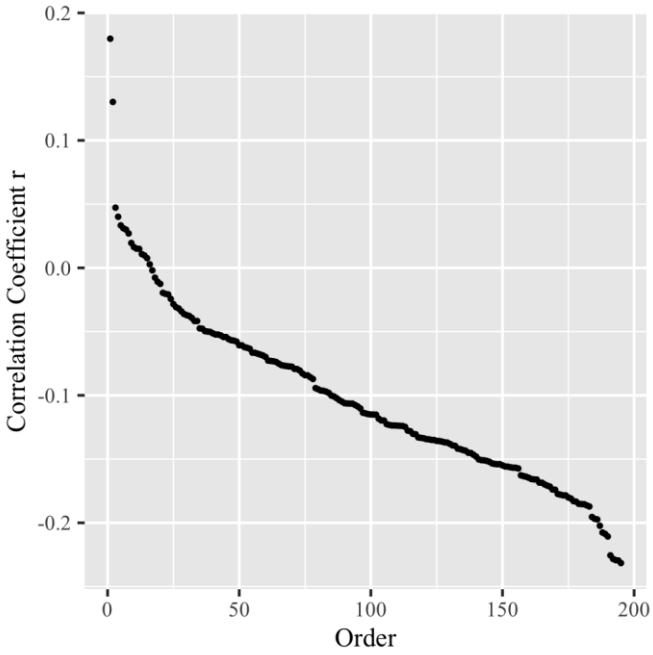

Figure 1. Correlation coefficients between freesurfer data and age ordered decreasingly according to value.

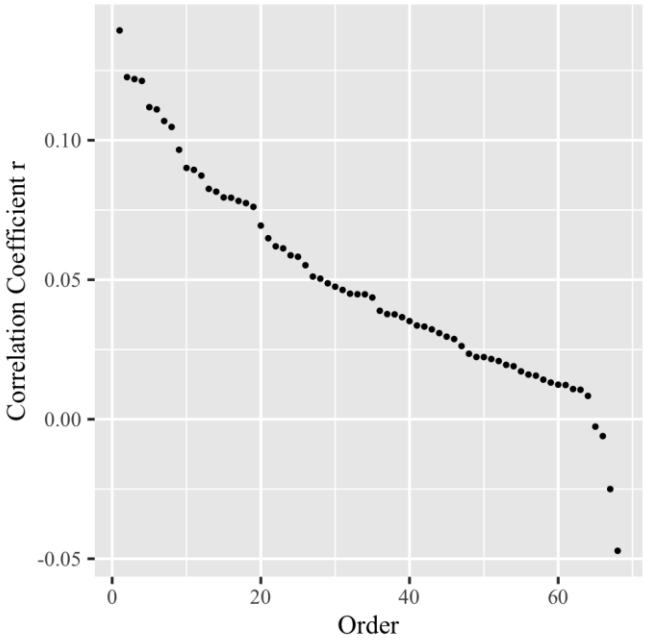

Figure 2. Correlation coefficients between isoperimetric-type ratios and age, ordered decreasingly according to value.

2.3 Bootstrap

In order to obtain estimates for the confidence intervals for the slope coefficients of the regression model, coefficient of determination, and accuracy of the multivariable (linear) regression model (MRM), a bootstrap was run 2000 times. In order to keep the process simple, the confidence interval utilized was 95 percent with the "percentile" interval being chosen due to how it appeared to do the best job describing the data given its distribution. The confidence interval of the coefficient of correlation is (0.1491,0.2265), and the confidence interval for the accuracy of the multivariable regression model is (0.4345,0.4718).

2.4 Prediction models

We used an MRM with the FreeSurfer data and also on the ITRs. Because our data is categorical, this approach did not work properly. Instead, we utilized multinomial logistic model (MLM) to compare the predictability of the ITRs to the untransformed FreeSurfer data. We ran models on the volume, thickness, and area and compared them to the model on the ITRs and volume to see if this dimensionality reduction had any effect on the adjusted R^2value. We did the same with models of thickness and area and compared them to models using the ITRs alone. We also created a model that combined the ITRs with the full FreeSurfer data. All of these attempted to predict all four age groups. Lastly, we created a model to predict whether the patient was in the age range of 22-30 or 31+, using only the ITRs as predictors. The MLMs were evaluated using a validation set approach with 10,000 iterations.

3. Results

Compared to the multivariable regression model with area and thickness data, the multivariable regression model trained on the ITRs had significantly lower prediction accuracy (table 1). As a result, we utilized MLM instead.

Table 1

Multivariable Regression Model Statistics

Predictors	Number of Variables	Age groups	R^2	$R^2{}_{adj}$
Thickness, Area, Volume	192	22-25, 26-30, 31-35, 36+	0.4393	0.3223
Isoperimetric	68	22-25, 26-30, 31-35, 36+	0.1271	0.0703

The MLM trained on the ITRs preserved the prediction accuracy of the MLM trained on the thickness and area data, with an overall increase of 0.018 in accuracy (table 2). Accordingly, the MLM trained on the ratios and volume data yielded similar results to the MLM trained on the FreeSurfer data.

Table 2

Multinomial Logistic Model Statistics

Predictors	Number of Variables	Age groups	Accuracy
Thickness, Area, Volume (FreeSurfer)	192	22-25, 26-30, 31-35, 36+	0.4196647
Thickness, Area, Volume, Isoperimetric	260	22-25, 26-30, 31-35, 36+	0.4199475
Thickness, Area	136	22-25, 26-30, 31-35, 36+	0.3916702
Isoperimetric	68	22-25, 26-30, 31-35, 36+	0.3987172
Isoperimetric and volume	124	22-25, 26-30, 31-35, 36+	0.4197851
Isoperimetric	68	22-30, 31+	0.618523

The quartile confusion matrix for the isoperimetric MLM shows that the model did not predict any subjects as being in the 36+ age group (table 3). Therefore, we decided to combine this age group with the 31-35 age group for future models.

Table 3

Quartile confusion matrix for isoperimetric-type ratio MLM

	22-25			26-30			31-35			36+		
	25%	50%	75%	25%	50%	75%	25%	50%	75%	25%	50%	75%
22-25	0.000	0.025	0.048	0.689	0.729	0.760	0.206	0.240	0.270	0.000	0.000	0.000
26-30	0.031	0.043	0.062	0.607	0.640	0.673	0.279	0.312	0.347	0.000	0.000	0.000
31-35	0.013	0.026	0.044	0.605	0.645	0.674	0.303	0.329	0.355	0.000	0.000	0.000
36+	0.000	0.000	0.000	0.500	0.500	1.000	0.000	0.500	0.500	0.000	0.000	0.000

4. Limitations

The ITRs are relatively nonlinear in relation to age. We attempted to use a multivariable regression model with transformations, but that did not yield a prediction accuracy similar to the FreeSurfer multivariable regression model. Thus, we used a MLM trained on the ITRs, which did yield consistent prediction accuracy with respect to the MLM trained on the FreeSurfer data.

The model's accuracy is limited by the behavioral dataset not containing all useful variables. The dataset does not contain perimeter, thickness, curvature, etc. which would have allowed us to estimate gyrification along with other morphometric changes that occur as one ages. We also could not calculate the ITRs for all 1207 subjects, because only 1113 subjects had 3T MR structural scans; therefore, we had to remove these subjects.

The model could be improved by using the complete dataset and exploiting the groups being ordered. We only used for the combined model (FreeeSurfer and isoperimetric) 192 of the approximately 575 variables in the behavioral dataset. As well, we could have taken advantage of the age groups being ordered by using a different model than the MLM that utilizes this feature.

5. Conclusions

As a result of the distribution of the correlations between the FreeSurfer data and age, we decided to use ITR as a dimensionality reduction technique. The ratio, as well, creates a more explainable model, where the ratio can be thought of as how the brain stretches.

The MRM trained on the ITRs failed to preserve the accuracy of the MRM trained on the area and thickness data. Therefore, we used a multinomial logistic model.

The MLM trained on the ITRs yielded a 50% dimensionality reduction and increased accuracy when compared to the MLM trained on the thickness and area variables. When the thickness and area was replaced with the ratios in the FreeSurfer model, the accuracy was preserved as well, with a 35.4% reduction in the number of variables. Therefore, the ratios are an explainable dimensionality reduction technique for modeling age using brain biometrics.

In the future, we plan on using other modeling techniques to exploit features in the dataset unused in our study. These features include the ages being ordered and the approximately 280 unused variables. We also plan on combining the 31-35 age group with the 36+ age group due to the small number of subjects in the 36+ age group.

6. References and citations

Al Zoubi, O., Ki Wong, C., Kuplicki, R. T., Yeh, H. W., Mayeli, A., Refai, H., … Bodurka, J. (2018). Predicting Age From Brain EEG Signals-A Machine Learning Approach. Frontiers in Aging Neuroscience, 10, 184. doi: 10.3389/fnagi.2018.00184

Aycheh, H. M., Seong, J.-K.., Shin, J.-H., Na, D. L., Kang, B., Seo, S. W., & Sohn, K.-A. (2018). Biological Brain Age Prediction Using Cortical Thickness Data: A Large Scale Cohort Study. Frontiers in Aging Neuroscience, 10, 252. doi: 10.3389/fnagi.2018.00252

Human Connectome Project. (2018). Behavioral Data [Data file]. Retrieved from https://www.humanconnectome.org/study/hcp-young-adult

P. Sturmfels, S. Rutherford, M. Angstadt, M. Peterson, C. Sripada, and J. Wiens. A domain guided cnn architecture for predicting age from structural brain images. arXiv preprint arXiv:1808.04362, 2018.

7. Appendix

7.1 Area, thickness, and volume data regions

Table 4

Brain regions for which the volume data is available

Left Lateral Ventricle	Right Lateral Ventricle	3rd Ventricle
Left Inferior horn of the Lateral Ventricle	Right Inferior horn of the Lateral Ventricle	4th Ventricle
Left Cerebellum White Matter	Right Cerebellum White Matter	Brain Stem
Left Cerebellum Cortex	Right Cerebellum Cortex	Cerebrospinal Fluid
Left Thalamus Proper	Right Thalamus Proper	5th Ventricle
Left Caudate	Right Caudate	WM Hypointensities
Left Putamen	Right Putamen	Optic Chiasm
Left Pallidum	Right Pallidum	Posterior Cingulate Cortex (CC)
Left Hippocampus	Right Hippocampus	Mid Posterior CC
Left Amygdala	Right Amygdala	Central CC
Left Accumbens-area	Right Accumbens-area	Mid Anterior CC

Left VentralDC	Right VentralDC	Anterior CC
Left Vessel	Right Vessel	
Left Choroid Plexus	Right Choroid Plexus	
Left WM (White Matter) Hypointensities	Right WM Hypointensities	
Lef non-WM Hypointensities	Right non-WM Hypointensities	

Table 5

Brain regions for which the thickness and area data are available

Left Banks of Superior Temporal Sulcus	Right Banks of Superior Temporal Sulcus
Left Caudal Anterior Cingulate	Right Caudal Anterior Cingulate
Left Caudal Middle Frontal	Right Caudal Middle Frontal
Left Cuneus	Right Cuneus
Left Entorhinal	Right Entorhinal
Left Fusiform	Right Fusiform
Left Inferior Parietal	Right Inferior Parietal
Left Inferior Temporal	Right Inferior Temporal
Left Isthmus Cingulate	Right Isthmus Cingulate
Left Lateral Occipital	Right Lateral Occipital
Left Lateral Orbitofrontal	Right Lateral Orbitofrontal

Left Lingual	Right Lingual
Left Medial Orbitofrontal	Right Medial Orbitofrontal
Left Middle Temporal	Right Middle Temporal
Left Parahippocampal	Right Parahippocampal
Left Paracentral	Right Paracentral
Left Pars Opercularis	Right Pars Opercularis
Left Parsorbitalis	Right Parsorbitalis
Left Pars Triangularis	Right Pars Triangularis
Left Pericalcarine	Right Pericalcarine
Left Postcentral	Right Postcentral
Left Posterior Cingulate	Right Posterior Cingulate
Left Precentral	Right Precentral
Left Precuneus	Right Precuneus
Left Rostral Anterior Cingulate	Right Rostral Anterior Cingulate
Left Rostral Middle Frontal	Right Rostral Middle Frontal
Left Superior Frontal	Right Superior Frontal
Left Superior Parietal	Right Superior Parietal
Left Superior Temporal	Right Superior Temporal

Left Supramarginal	Right Supramarginal
Left Frontal Pole	Right Frontal Pole
Left Temporal Pole	Right Temporal Pole
Left Transverse Temporal	Right Transverse Temporal
Left Insula	Right Insula