Volumetric Normalisation of Anatomical Brain Structures
Final Report

Greenland Yu
The University of Adelaide
a1740184@student.adelaide.edu.au

Mark Jenkinson - Supervisor
Mark.Jenkinson@adelaide.edu.au

Stephan Lau - Supervisor
Stephan.Lau@adelaide.edu.au

Abstract

Volumetric normalisation of brain structures needs to be performed to obtain comparable volumes between patients. Currently there exists no study that investigates adjustment methods of brain structures using a large dataset. In this study, different adjustment methods are analysed using a recently available dataset, the UK BioBank, to determine the effectiveness of current methods. To complete this, a review into the current literature of adjustment methods is done to collect the most promising adjustment methods. These adjustment methods were the proportion approach, the power-proportion approach and the general linear model approach. Next these adjustment methods were applied to the UK BioBank so that their effectiveness can be evaluated. It was found that the power-proportion and general linear model approach out performed the proportion approach in all tests. Meaning that the proportion approach should no longer be used. The power-proportion and general linear model approach performed similarly well on all tests, the definitive choice between which of the two adjustment methods should be used needs further investigation. The caveat to the general linear model approach is that it assumes confounding factors have a linear relationship with the volume of interest. By disregarding the proportion approach and using either the power-proportion and general linear model approach, neurologists can make better informed decisions on patient treatment.

1. Introduction

In the field of neuroimaging, structural imaging focuses on analysing the structure of the nervous system and diagnosing neurological diseases such as tumours. To do so, calculating accurate and comparable intracranial volumes is often crucial for correct diagnosis. Individuals vary in head size and this has a natural effect on brain volumes; a larger head size associates to more brain volume. Due to this, head size must be taken into consideration when comparing different individuals' brain volumes. Head size or more precisely total intracranial volume acts as a confounding factor, raw brain volumes must be normalised before comparison. The purpose is that when diagnosing neurological diseases, neurologists can compare a patient's normalised brain volumes to a standard set of brain volumes to see if there are any signs of atrophy. Currently, the studies for evaluating the effectiveness of different methods for normalising brain volumes use small datasets to draw statistical conclusions from. This project strives to use the recently available UK BioBank containing 37,853 individuals to evaluate these different normalisation/adjustment methods.

2. Literature Review

One of the earlier approaches to normalising brain volumes is by using total intracranial volume (TIV) as an adjustment factor and by simply dividing the brain volume of interest (VOI) by total intracranial volume. This is the proportion approach:

$$VOI_{\text{scaled}} = \frac{VOI}{TIV} \tag{1}$$

This was documented in the 2001 paper by Whitwell et al. [1] in which they conducted a study (n=67) to show that normalisation with TIV and the proportion approach reduces inter-individual variation. Their results show the coefficient of variation was reduced between individuals in their dataset from 10% to 6% where $P<0.001$, showing that the result is highly significant. Limitations introduced in the data collection stage include movement artifacts, machine-dependent image to image variation and voxel size variations due to drifts in scanner calibration. Another limitation is the small sample size used to conduct this study. With only 67 participants, any statistical conclusions drawn do not have strong backing. An expansion on this study would be to use the proportion approach on a larger dataset to verify their claims of reducing inter-individual variation. Another extension would be to test whether individual brain structures can be normalised using the same methods as total brain volume. Brain structures are specific parts of the brain, such as the hippocampus and thalamus. Doing so would reveal insight into whether brain structures scale "globally" or in a "local" manner.

O'Brien et al. [2] conducted a review on common adjustment factors and normalisation methods. For each adjustment factor and statistical method the authors provided an in-depth discussion.

Using head circumference as an adjustment factor was found to work best for children 6 years of age and younger (R=0.93), weak for 7 to 16 years of age (R=0.67) and also weak for ages 17 to 42 (R=0.69). Consequently, head circumference should only be seriously considered in circumstances where the patient is 6 years of age or under. The second type of adjustment factor investigated was body parameters, which include height and weight. Two studies [3][4] were presented where the researchers came to contradicting conclusions on the usefulness of body parameters as a way to normalise brain volume. Therefore, one cannot assume that body parameters hold a strong or consistent relationship to brain volumes for all demographics and therefore should not be used as an adjustment factor. The last adjustment factors analysed were volumetric factors, these were total brain volume (TBV) and total intracranial volume (TIV). The authors stated that compared to TBV, TIV is a preferred adjustment factor in subjects of older age. This is due to TBV decreasing in adulthood and onwards, and we cannot assume that all volumes of interest decrease at the same rate.

The proportion normalisation method is also mentioned by O'Brien et al. Contrasting Whitwell et al [1], O'Brien states that brain volume measures are typically used as the adjustment factor, whereas Whitwell et al [1] exclusively used the total intracranial volume as the adjustment factor. The next normalisation method mentioned in O'Brien is the general linear model, or more specifically the analysis of covariance (ANCOVA) approach. This is where the data is fitted in groups to a regression model. The regression model is shown here:

$$\boldsymbol{Y} = \boldsymbol{B} * (\boldsymbol{X} - \overline{x}) + \mu + \boldsymbol{\varepsilon} \tag{2}$$

Where the dependent variable $\boldsymbol{Y}$ is a matrix corresponding to the VOI. The covariate $\boldsymbol{X}$ is also a matrix, covariates includes the adjustment factor as well as other group indicators such as physiological state and age. $\overline{x}$ is the global mean for covariate x and μ is the grand mean of the covariates. Variables to be fitted are $\boldsymbol{B}$ which are the slopes, in the form of a vector of size n where n is the number of covariates. $\boldsymbol{\varepsilon}$ represents the unobserved error term. A drawback of the ANCOVA approach is that it assumes a linear relationship between the dependent variable (VOI) and the continuous covariate(s), where in real life this relationship may not be linear.

Another regression based volume analysis method is the residual approach. O'brien et al. describes this as where data from the control group is taken and linear regression applied to it with the volume of interest (VOI) as the target output value and relevant covariates as the training set. Next this regression model is applied to the rest of the dataset (diagnosis group) and the outputted predicted values are subtracted from the observed volume. This difference is defined as the residuals. The residuals represent the deviation of an individual's VOI to that of a control. To determine the significance of deviation the diagnosis group has from the control group, analysis of variance or a t-test can be carried out. A constraint that the residual approach has is that it can only be used on data that is separated into two groups, the control group and the diagnosis group. This method is not an adjustment method but rather a statistical measure of differences between the brain volumes of two groups.

Liu et al. [5] argues that both the proportionality approach and the analysis of covariance approach assumes that the volume of interest is proportionally or linearly related to TIV, where in real life this may not be the case. Instead Liu et al. proposes a TIV correction method that describes brain structures as having a power relationship with TIV. More formally as:

$$VOI = \alpha * TIV^{\beta} \tag{3}$$

Where VOI is the volume of interest, α is a constant, TIV is total intracranial volume (the adjustment factor) and β is the scaling exponent of the power function. The power-proportion approach finds an estimate of β by fitting Model 3 with sample data, this estimate is denoted as b. Then normalisation of new data can be done by applying the following formula:

$$VOI_{PPC} = \frac{VOI}{TIV^{b}} \tag{4}$$

Where VOI_{PPC} is the power-proportion corrected volume of interest. To show that the power-proportion method normalised volumes by getting rid of the effect of different sized TIV, a scatter-plot of VOI_{PPC} (dependent variable) against TIV (independent variable) was plotted. Additionally, a scatter-plot of proportion adjusted volumes against TIV was also plotted for comparison. Liu et al. reasons that the since the linear regression line produced by the power-proportion plot is more parallel to the x-axis compared to the proportion adjusted plot; it showed the power-proportion approach removed the relationship between the adjusted volume of interest and TIV more effectively than the proportion approach. Furthermore, leave-one-out cross-validation was conducted to compare between the power-proportion method and the ANCOVA method, the results showed that the ANCOVA method had an overall higher error rate. It was noted that ANCOVA results only had high error rates for brain structures that had fitted β values significantly different than 1. When the β value for structures were close to 1, the error rates for ANCOVA and power-proportion adjustment were similar. Liu et al. comments that "this is not surprising as the power-proportion method includes ANCOVA method as a special case". A limitation of this paper was the small sample size used to conduct the study, with only 141 samples. An extension to this study would be using a differently sourced dataset and performing the power-proportion fitting to compare β values.

3. Methodology

As this project is more of an exploratory typed project, the methodology was designed so that results were produced at each stage and further investigations could be done at any point.

The first step was to conduct a literature review exploring the current normalisation methods, and through this note which methods are the most promising and should be investigated. As mentioned in the above literature review, the three most promising normalisation methods were the proportion approach, the power-proportion approach and the general linear model approach.

3.1. Data Source

The data used in this project was sourced from the UK BioBank. The BioBank, provides a wealth of health related data. For this project, data related to brain structure volumes and patient features were used. Patient features included sex, age and BMI.

The UK BioBank data was explored to gain insight into the magnitude and ranges of the volumetric structures. This was done visually through plotting histograms of the data and quantitatively through generating measures of mean, quartiles and spread of the data.

3.2. Data Preparation

Before meaningful adjustment of volumes, the dataset needed to be cleaned up. The data cleaning process can be split into three stages. Here they are listed along with the reasoning behind each stage.

1. Remove rows with missing entries. This is so that adjustments are not done on missing entires, producing meaningless data.

2. Remove outliers. Due to the sheer amount of data points in the dataset, we can confidently remove outliers.

3. Normalise the total intracranial volume column such that the mean value is 1. This is achieved by dividing all TIV values by the mean TIV value. The reasoning is that when performing the adjustment methods the adjusted volumes are in the same units of magnitude.

3.3. Implementation and Evaluation

Next the adjustment methods described in the literature review were implemented in Python so that data manipulation, data transformation and data analysis can all be automated. The sci-py stack was used to implement this framework.

The proportion approach was the easiest adjustment method to implement due to its inherent simplicity. The proportion adjustment function, accepted the whole dataset, the adjustment factor (usually TIV) index in the dataset and the VOI index in the dataset. Since every row represented an individual, for ever row, the proportion adjustment function divided the VOI value by the TIV value and returned a column of the adjusted volumes. This proportion adjustment function is wrapped inside a conduct proportion analysis function that performs this adjustment and then produces plots and statistical analysis of these volumes. More specifically, the analysis function created a density plot and a line of best fit within the density plot for the adjusted and unadjusted volumes, saving the plots in a PDF format. Additionally, the coefficient of variation and statistics produced using the pandas describe function was appended onto the PDF as extra information. The pandas describe function provides count, mean, standard deviation and quartile measures.

The power-proportion approach is a little bit more involved in its implementation. As described in the literature review, for every VOI a β value needed to be found through fitting a power law relationship between the VOI and TIV. After finding this β value for a particular VOI then adjustments can be made for that volume. To first fit $VOI = \alpha * TIV^{\beta}$ with sample data, it is helpful to transform this equation using log laws into one resembling a linear relationship so a linear regression can be performed.

$$
\begin{gathered}
VOI = \alpha * TIV^{\beta} \\
log(VOI) = log(\alpha * TIV^{\beta}) \\
log(VOI) = log(\alpha) + log(TIV^{\beta}) \\
log(VOI) = log(\alpha) + \beta * log(TIV)
\end{gathered}
$$

In this form, the β value is the slope of the linear relationship between the volumes that have had the logarithmic function applied. By using the sci-kit learn LinearRegression class to find the model that fits the transformed volumes the slope of that trained model is found. With this β value adjustments can then be done to the VOI. The power-proportion adjustment function divides the VOI by the TIV raised to the power of the β value. Similar to the proportion adjustment function, the power-proportion function returns a column of adjusted volumes and itself is wrapped in a conduct power-proportion analysis function that performs additional analysis on these adjusted volumes. This analysis is the same as the ones done for the proportion approach, with the added statistic of the fitted β value to the appended PDF.

Implementation of the general linear model adjustment method required some transformations to the data before fitting with the sci-kit learn library. The features that are to be used in the multiple linear regression were sex, BMI, age and the TIV. Before fitting, these features needed to be transformed so that each feature's mean is zero. The spread of the features should remain the same. This transformation is necessary because when fitting we do not want the mean of the VOI to change; by setting the mean of the features to 0, running a regression will not change the VOI mean.

After this transformation, the features and the VOI values were fed into the sci-kit learn LinearRegression class, which also performs multiple linear regression. The fitted model represents the relationship between the zero-centered confounding factors and the VOI. By using this model to predict for a VOI and subtracting the predictions from the original VOI, this final volume represents the zero-centered VOI with confounding factors removed. All that needs to be done is to add the mean of the original VOI to the zero-centered VOI. This whole process was implemented in a single function save for the feature adjustment which was done in a helper function. These adjusted volumes underwent the same analysis as the other adjusted volumes produced by the other adjustment methods.

4. Experimental Setup

In the process of generating the PDFs for each VOI and each adjustment method, the effectiveness of each adjustment method is also tested. A way of determining how well an adjustment method removed confounding factors is by plotting the adjusted VOI as the dependent variable and the confounding factor as the independent variable. If there were no relationship between the dependent and independent variable, the slope for the line of best fit would be close to zero.

A second way of determining the effect of an adjustment method is by looking at the spread of the volumes before and after adjustment. Either visually through plotting or quantitatively through calculating the coefficient of variation. The coefficient of variation is the standard deviation divided by the mean. By having a tighter spread, volumes are more comparable to one another.

The last way of evaluating each method's effectiveness is through looking at how well a linear classifier performed with different volumes as features. Using the sci-kit learn support vector machine, a model was trained to predict whether the patient was young (under 55) or old (above 70) using sex, BMI and a VOI as features. Adjusted volumes were expected to perform better since the volumes were less spread out.

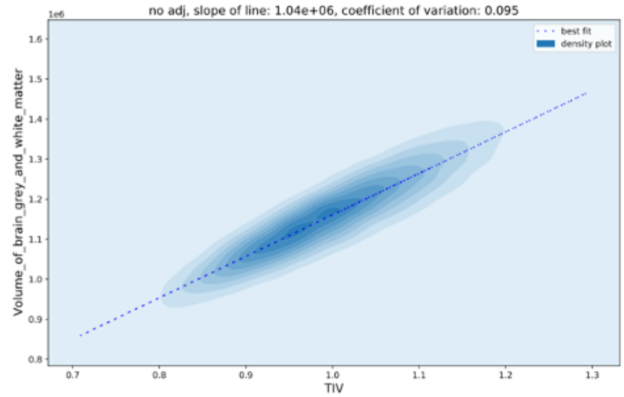

Figure 1. TIV vs TBV, unadjusted

5. Results

The dataset provided a plethora of different brain structures, each of these brain structures were adjusted using the three implemented adjustment methods. The different brain structures included total brain volume, average thalamus volume, average caudate volume, average putamen volume, average pallidum volume, average hippocampus volume, average amygdala volume and average accumbens volume. Average volume means the average of the left VOI and the right VOI. Due to the sheer amount of data points being plotted, data points would overlay each other making the density of regions unclear. This is why density plots have been chosen to visualise the data, darker regions in the plots signify a higher density of data points in that area. A line of best fit, the dotted line, is also drawn through the plots. The unadjusted volume is included in the table of slopes and table of coefficient of variation as a point of comparison to the adjusted volumes.

Table 1 describes the slopes of the linear regression line fitted to the VOI and TIV. The units of all volumes are in mL, so the slope is a unit-less measure. The slopes for the general linear model adjusted column has been converted to the same exponent in scientific notation (e-12), this is for ease of interpretation.

As described in the Experimental Setup section, the coefficient of variation of the VOI is another method of assessing the effectiveness of an adjustment method. Table 2 describes the coefficient of variation for each VOI and each adjustment method.

A classifier can also be used in testing the effectiveness of adjustment methods. As summarised in Table 3, accuracy measures were produced for only two VOI, the TBV and the hippocampus.

6. Discussion

The first Volume of Interest examined is the Total Brain Volume (TBV). As can be seen in Figure 1, TIV appears to

Volume	Unadjusted	Proportion adj.	Power-proportion adj.	GLM adj.
Accumbens	322	-115	10.2	8.65e-12
Amygdala	871	-367	8.91	26.1e-12
TBV	1040000	-123000	-144	2150e-12
Caudate	2440	-1030	-0.374	68.1e-12
Hippocampus	1810	-2010	35.3	83.9e-12
Pallidum	1140	-639	1.49	36.7e-12
Putamen	3410	-1380	5.43	97.4e-12
Thalamus	4990	-2640	9.65	156e-12

Table 1. Slopes of the linear regression lines of Volumes of Interest and Total Intracranial Volume for different adjustment methods.

Volume	Unadjusted	Proportion adj	Power-proportion adj.	GLM adj.
Accumbens	0.237	0.225	0.225	0.204
Amygdala	0.175	0.163	0.161	0.161
TBV	0.095	0.04	0.039	0.033
Caudate	0.122	0.105	0.101	0.101
Hippocampus	0.115	0.115	0.104	0.101
Pallidum	0.13	0.119	0.114	0.113
Putamen	0.12	0.102	0.098	0.092
Thalamus	0.097	0.08	0.073	0.067

Table 2. Coefficient of variation for Volumes of Interest, rounded to 3 decimal places.

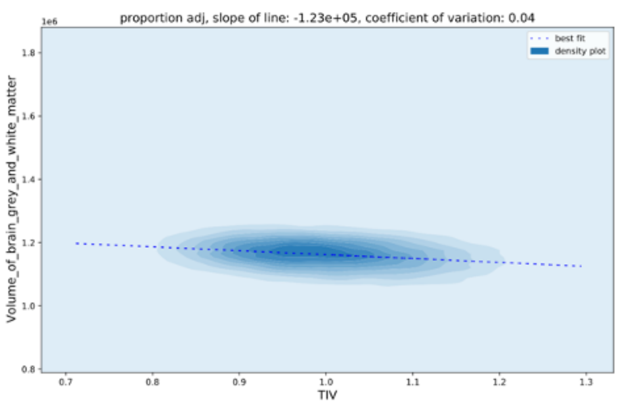

Figure 2. TIV vs TBV, proportion adjusted

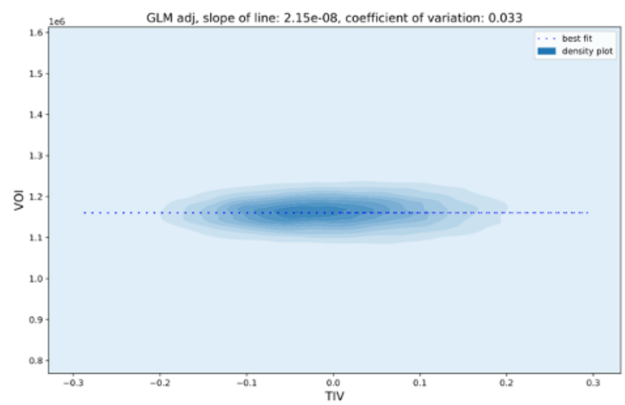

Figure 4. TIV vs TBV, general linear model adjusted

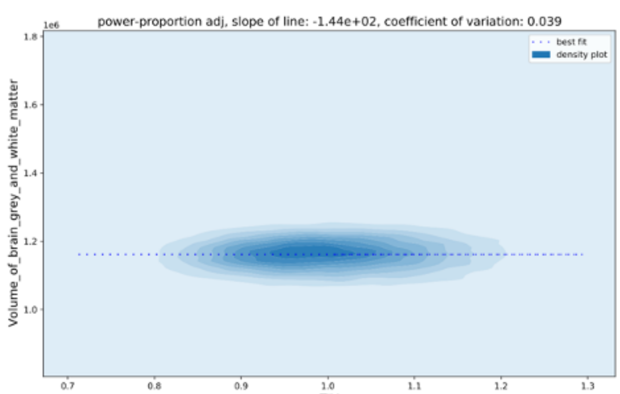

Figure 3. TIV vs TBV, power-proportion adjusted

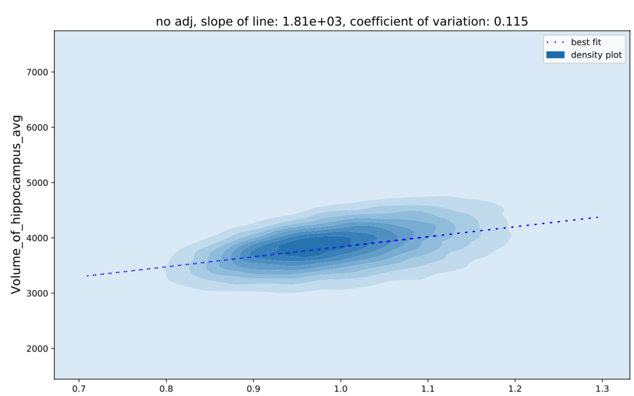

Figure 5. TIV vs hippocampus, unadjusted

Adjustment Method	Total Brain Volume Accuracy	Hippocampus Accuracy
Unadjusted	0.716	0.678
Proportion adj.	0.805	0.619
Power-proportion adj.	0.828	0.603
General Linear Model adj.	0.829	0.676

Table 3. The accuracy of a classifier predicting between two age ranges (whether an individual is below 55 or above 70) rounded to 3 decimal places. Volumes of interest analysed are total brain volume and hippocampus.

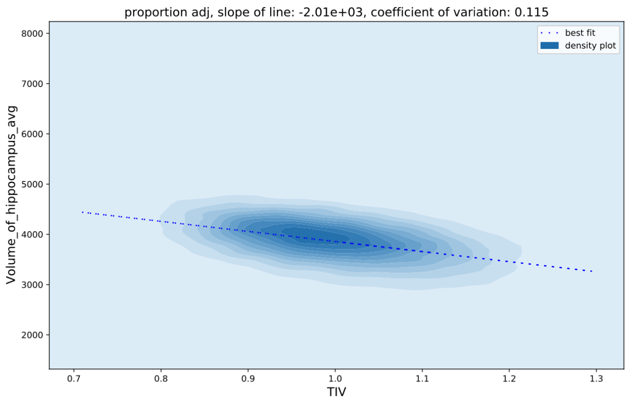

Figure 6. TIV vs hippocampus, proportion adjusted

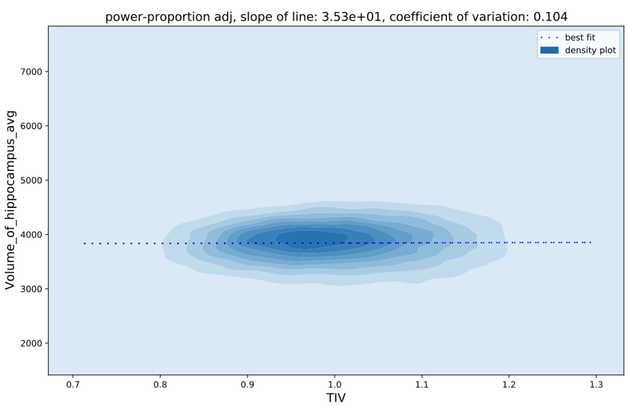

Figure 7. TIV vs hippocampus, power-proportion adjusted

have a proportional/linear relationship with TBV. For individuals with larger TIV, they also have larger TBV. When the proportion adjustment method is applied, seen in Figure 2, the TBV volumes for individuals with large TIV has been adjusted to about the same level as those with smaller TIVs. However, on closer inspection, the slope of the line of best fit is negative (at -123000), meaning that the adjusted TBV is negatively correlated to TIV. By applying the power-proportion adjustment method, seen in Figure 3, the line of best fit is parallel to the x-axis. This shows that after power-proportion adjustment, the relationship between TIV and TBV have become uncorrelated. Even though the slope of the power-proportion adjustment method is -144, given the magnitude of TBV it is not significantly non-parallel. The general linear model adjusted plot, Figure 4, appears to

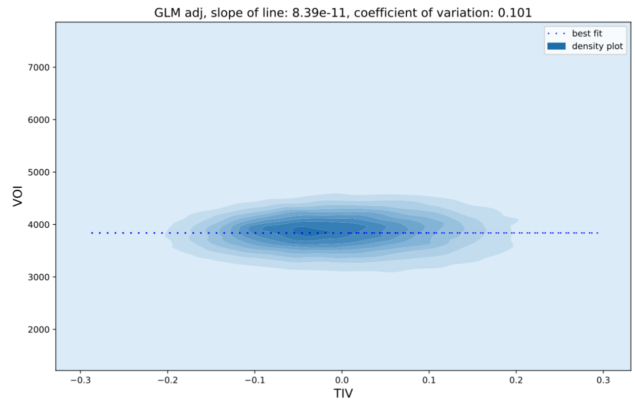

Figure 8. TIV vs hippocampus, general linear model adjusted

be very similar to Figure 3. Closer inspection reveals that the slope of the general linear model adjusted plot is near 0. This slope of near 0 is to be expected as the general linear model finds the linear relationship between a VOI and TIV and removes that relationship when adjusting.

The coefficient of variation for unadjusted TBV is 0.095, after any sort of adjustment this number decreases. For proportion adjustment it decreases to 0.04, 0.039 for power-proportion adjustment and 0.033 for the general linear model adjustment. This can also be visually deduced from the plots as the density contours are more tightly packed to the line of best fit.

Another measure of adjustment effectiveness was the classifier. When any of the adjustment methods were used the classifier accuracy rose. This showed that adjusting TBV is beneficial when the goal is to classify patient traits.

The other Volume of Interest discussed is the hippocampus, chosen because of its major association with learning and memory[8]. As seen in Figure 5, the unadjusted hippocampus volume has a positive relationship with TIV. When the proportion adjustment method is applied, seen in Figure 6, this positive relationship becomes negative. Looking at the slopes in Table 1, the proportion adjustment method slope in fact was even less parallel than the unadjusted slope. This indicates that when the poportion adjustment method is applied to the hippocampus volume it does not remove the confounding effect of TIV, in fact it exasperates it. When the power-proportion adjustment

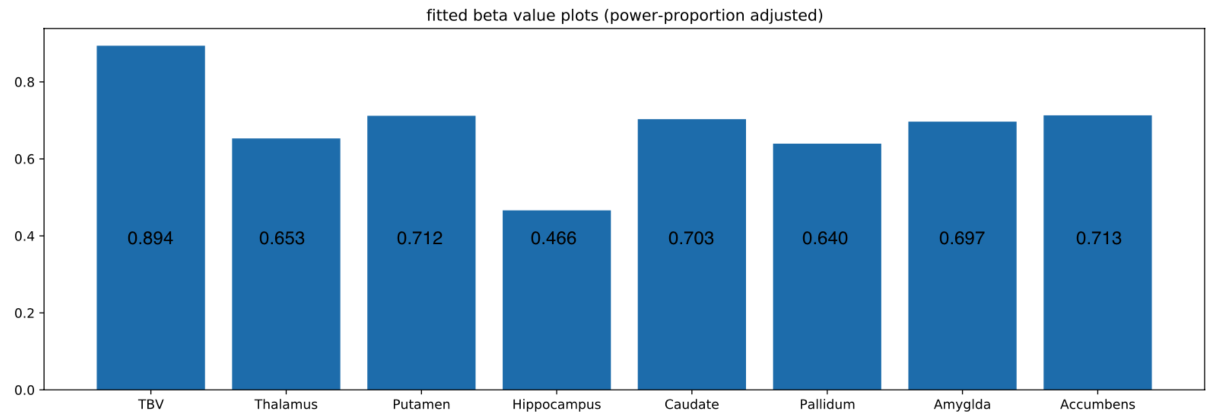

Figure 9. The beta values for each VOI, arranged in descending VOI volume

method is applied, seen in Figure 7, the slope of the line of best fit is much more parallel to the x-axis. This shows the power-proportion method has removed the confounding effect of TIV. Lastly, the general linear model adjusted plot, Figure 8, appears very similar to the plot of the power-proportion adjusted volumes. On a closer inspection, the general linear model adjustment slope is very near 0 while the power-proportion slope is 35.3. This is to be expected as the general linear model adjustment completely removes the relationship between a VOI and covariates. Considering the magnitude of the hippocampus volume, the power-proportion slope is not significantly different from 0, so it is comparable to the general linear model adjusted slope.

The adjusted hippocampus volumes differ from the adjusted TBV volumes in that the coefficient of variation did not significantly lessen after adjustment. The coefficient of variation when the proportion approach was applied was the same for unadjusted hippocampus volumes, showing that the proportion approach did nothing to decrease the spread of volumes. The coefficient of variation did reduce when the power-proportion and general linear model adjustment method was applied but not significantly.

The classifier accuracy in fact decreased for all adjustment methods when the VOI was hippocampus. With the exception of the general linear model adjustment method which only decreased slightly, from 67.8% accuracy to 67.6% accuracy. This is in contrast to the classifier accuracy of TBV where all adjusted volumes performed better than the unadjusted volume.

As seen in Figure 9, the fitted beta values when performing the power-proportion adjustment method are all significantly different from 1. If the fitted beta value was 1, the power-proportion approach would be performing the proportion adjustment method. This result contrasts Liu et al [5] where that study had multiple volumes with beta values being not significantly different from 1. All of the fitted beta values in this study were lower than the ones found in the Liu [5] study, this showed that the VOI for this study had a more significant power relationship.

A limitation to this study is that the classifier, used to classify brains as young or old, was not cross-validated to have optimised hyper-parameters. The classifier used was a support vector machine from the scikit-learn library with kernel type 'poly' and degree 3. Default hyper-parameters were used in the training stage. This means that the accuracy of classification will not be the highest possible. Nevertheless, the results still have meaning to it and can be used as a benchmark when comparing adjustment methods. An extension of this study would be to properly cross-validate the model and use optimised hyper-parameters to generate the prediction accuracy. Another limitation to this study is that the individuals in the dataset were all assumed to be healthy and have no neurodegenerative disease that would effect brain volume. This would not allow for analysis into how the different adjustment methods correct for when brain structures have been atrophied. An extension to this study would be to extract atrophied individuals from the UK BioBank and perform analysis on those data points as well.

7. Conclusion

By using data from the UK BioBank it has been determined that while historically used, the proportion adjustment method does not remove confounding variables as well as the power-proportion and general linear model adjustment methods. This is evident from the plots of proportion adjusted volumes still having a relationship with TIV. The power-proportion and general linear model adjustment methods perform similarly in terms of removing confounding variables. However, the general linear model outperforms the power-proportion method in reducing the spread of volumes. This provides room for further investigation into different evaluation methods for these adjustment factors.

References

[1] Whitwell, J., 2001. Normalization of Cerebral Volumes by Use of Intracranial Volume: Implications for Longitudinal Quantitative MR Imaging. American Society of Neuroradiology, 22. 1, 2

[2] OBrien, L., Ziegler, D., Deutsch, C., Kennedy, D., Goldstein, J., Seidman, L., Hodge, S., Makris, N., Caviness, V., Frazier, J. and Herbert, M., 2006. Adjustment for Whole Brain and Cranial Size in Volumetric Brain Studies: A Review of Common Adjustment Factors and Statistical Methods. Harvard Review of Psychiatry, 14(3), pp.141-151. 2

[3] Passingham, R., 1979. Brain Size and Intelligence in Man. Brain, Behavior and Evolution, 16(4), pp.253-270. 2

[4] Jerison HJ. The evolution of diversity in brain size. In: Hahn ME,Jensen C, Dudek BC, eds. Development and evolution in brain size. New York: Academic, 1979:29–57. 2

[5] Liu, D., Johnson, H., Long, J., Magnotta, V. and Paulsen, J., 2014. The power-proportion method for intracranial volume correction in volumetric imaging analysis. Frontiers in Neuroscience, 8. 2, 7

[6] Géron, A., n.d. Hands-On Machine Learning With Scikit-Learn And Tensorflow. 2nd ed.

[7] Hansen, T., Brezova, V., Eikenes, L., Håberg, A. and Vangberg, T., 2015. How Does the Accuracy of Intracranial Volume Measurements Affect Normalized Brain Volumes? Sample Size Estimates Based on 966 Subjects from the HUNT MRI Cohort. American Journal of Neuroradiology, 36(8), pp.1450-1456.

[8] Dhikav, V. and Anand, K., 2012. Hippocampus in health and disease: An overview. Annals of Indian Academy of Neurology, 15(4), p.239. 6