Neuroimage: Reports 2 (2022) 100137

Contents lists available at ScienceDirect

Neuroimage: Reports

journal homepage: www.sciencedirect.com/journal/neuroimage-reports

Cortical thickness in brain imaging studies using FreeSurfer and CAT12: A matter of reproducibility

Maria de Fátima Machado Dias [a,b,*], Paulo Carvalho [b,d], Miguel Castelo-Branco [a,c,d], João Valente Duarte [a,c]

[a] CIBIT, ICNAS, University of Coimbra, Portugal
[b] CISUC, Faculty of Sciences and Technology, University of Coimbra, Portugal
[c] Faculty of Medicine, University of Coimbra, Portugal
[d] Intelligent Systems Associate Laboratory (LASI), Portugal

A B S T R A C T

A reproducibility crisis has been reported across many research fields, including neuroimaging, reaching up to 70% of studies. Neuroimaging data, such as magnetic resonance imaging (MRI), requires pre-processing to allow for inter-subject comparison, increase signal contrast and noise reduction. As manual MRI pre-processing is time consuming and requires expertise, multiple automatic pre-processing frameworks have been proposed. However, neuroimaging studies often report divergent results, even for similar populations, thus it is important to determine whether this occurs as a result of different processing tools. Two of the most used tools are FreeSurfer and the Computational Anatomy Toolbox (CAT12). In this study we assessed the reproducibility between these two automatic pre-processing frameworks for structural MRI and test-retest reliability within framework on estimation of cortical thickness. Our results show that the reproducibility between the frameworks is lower at the region-of-interest (ROI) level than at individual level. Furthermore, we found that the reproducibility was lower in paediatric samples than in adults. Finally, an acquisition site effect was also identified. Given the widespread use of these frameworks in basic and clinical neuroscience, the results of multicentric cross-sectional studies must be interpreted with caution, particularly with paediatric samples. The observed reproducibility issue might be one of the sources of discrepancies reported in neuroimaging studies. On a positive note, framework test-retest reliability within subject is high, suggesting that inconsistency of results may be less concerning in longitudinal studies. The code is available at: https://cibit-uc.github.io/fs-cat12-cortical-thickness-reproducibility.

1. Introduction

Structural magnetic resonance imaging (MRI) allows to estimate various morphometric features of the brain such as cortical thickness, which is widely used in basic and clinical research. The cortical thickness of healthy brains in humans measures, on average, 2.5 mm (Hutton et al., 2008), and is associated with cognitive function (Dickerson et al., 2008; KL et al., 2007). Cortical thickness is reported to be correlated with disease progression in Alzheimer's disease (Dickerson et al., 2008; Ossenkoppele et al., 2019), schizophrenia (Van Haren et al., 2011; Yan et al., 2019), and Parkinson's disease (Burton et al., 2004; Ibarretxe--Bilbao et al., 2012; Jubault et al., 2011). Nonetheless, inconsistent results have been reported regarding cortical thickness changes in neurodevelopment (Walhovd et al., 2016) and in multiple diseases such as Autism Spectrum Disorder (ASD) (Bedford et al., 2020; Khundrakpam et al., 2017; Van Rooij et al., 2018; Wallace et al., 2010) bipolar disorder (Hibar et al., 2018), or major depression disorder (Li et al., 2019).

Cortical thickness corresponds to the distance between the pial surface of the brain and the white matter boundary. Despite this straightforward definition, the evaluation of the cortical thickness is difficult given the inherent challenge of creating accurate thickness estimations due to the highly folded brain morphology. The manual pre-processing of MRI images is time consuming and requires field expertise, therefore different computational frameworks have been proposed to automatically process MR images and to compute morphometric measures (Ashburner and Friston, 2005; Dora et al., 2017; Fischl, 2012; Vovk et al., 2007). The automation has several advantages: it does not require deep user expertise, it can be parallelized, errors are systematic rather than random, and its reproducibility is amenable to explicit testing. Yet, currently, there is no gold standard pre-processing framework. Moreover, each framework uses its implementation to pre-process MRI data and extract brain tissue estimations such as cortical thickness. The different implementations can be a source of variability, which might explain inconsistencies reported in literature, as mentioned above.

This study assesses the cortical thickness estimation's reproducibility

* Corresponding author. CIBIT, ICNAS, University of Coimbra, Portugal.
E-mail address: maria.dias@uc.pt (M.F.M. Dias).

https://doi.org/10.1016/j.ynirp.2022.100137
Received 24 March 2022; Received in revised form 14 September 2022; Accepted 16 September 2022
Available online 7 October 2022
2666-9560/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

between two of the most widely used pre-processing frameworks: FreeSurfer and Computational Anatomy Toolbox (CAT12). The cortical thickness estimations' reproducibility between CAT12 and FreeSurfer was evaluated previously in very small sample size studies (Seiger et al., 2018; Velázquez et al., 2021), in which authors reported a mean R^2 and correlation between frameworks, for cortical thickness, of 0.89 and 0.92, respectively. However, the studies exhibit multiple limitations: the number of subjects used in the study was very low; the studies only focused on an adult population, and therefore did not consider the changes that occur in early life periods such as childhood and adolescence; and the studies did not assess whether the reproducibility depends upon the site of acquisition setting and individual age. Another study (Masouleh et al., 2020) also evaluated the pre-processing pipeline impact on cortical thickness estimation. Nevertheless, the study did not assess the reproducibility at the subject level. To overcome these shortcomings, the current work uses MRI data from three open data sharing initiatives with a large number of subjects. Furthermore, our study investigates whether the reproducibility between the two frameworks depends upon the acquisition setting, the subjects' age and brain region (region-of-interest, ROI). In particular, the objectives of the current work are (a) to compare the test-retest reliability within each framework, (b) to investigate whether the reproducibility of FreeSurfer and CAT12 frameworks depends on the age and neurodevelopmental stage, and (c) to assess these frameworks' reproducibility on different acquisition settings and verify whether different frameworks yield different results modelling age using cortical thickness values.

2. Material and methods

This work evaluates the reliability of cortical thickness estimations within two frameworks: CAT12 and FreeSurfer, and the reproducibility of cortical thickness estimations between these two frameworks. The study was divided into three objectives:

- Objective (a) assesses and compares the test-retest reliability of each framework (FreeSurfer and CAT12). In this objective, we analysed two T_1-weighted images (Run_1 and Run_2) from each participant and cortical thickness is extracted for both using both frameworks. The reliability metrics are computed within framework using the cortical thickness estimations of the two images. Therefore, two groups were considered in this objective and the participants were the same in both groups. For this analysis the OASIS-3 dataset was used, including all participants who had two consecutive images acquired within less than 60 min.
- Objective (b) compares the reproducibility metrics between two age-groups, the paediatrics group i.e., individuals with age lower or equal than 18 years old, and the adults group, i.e., individuals with age greater than 18 years old. For this objective the ABIDE I dataset was used.

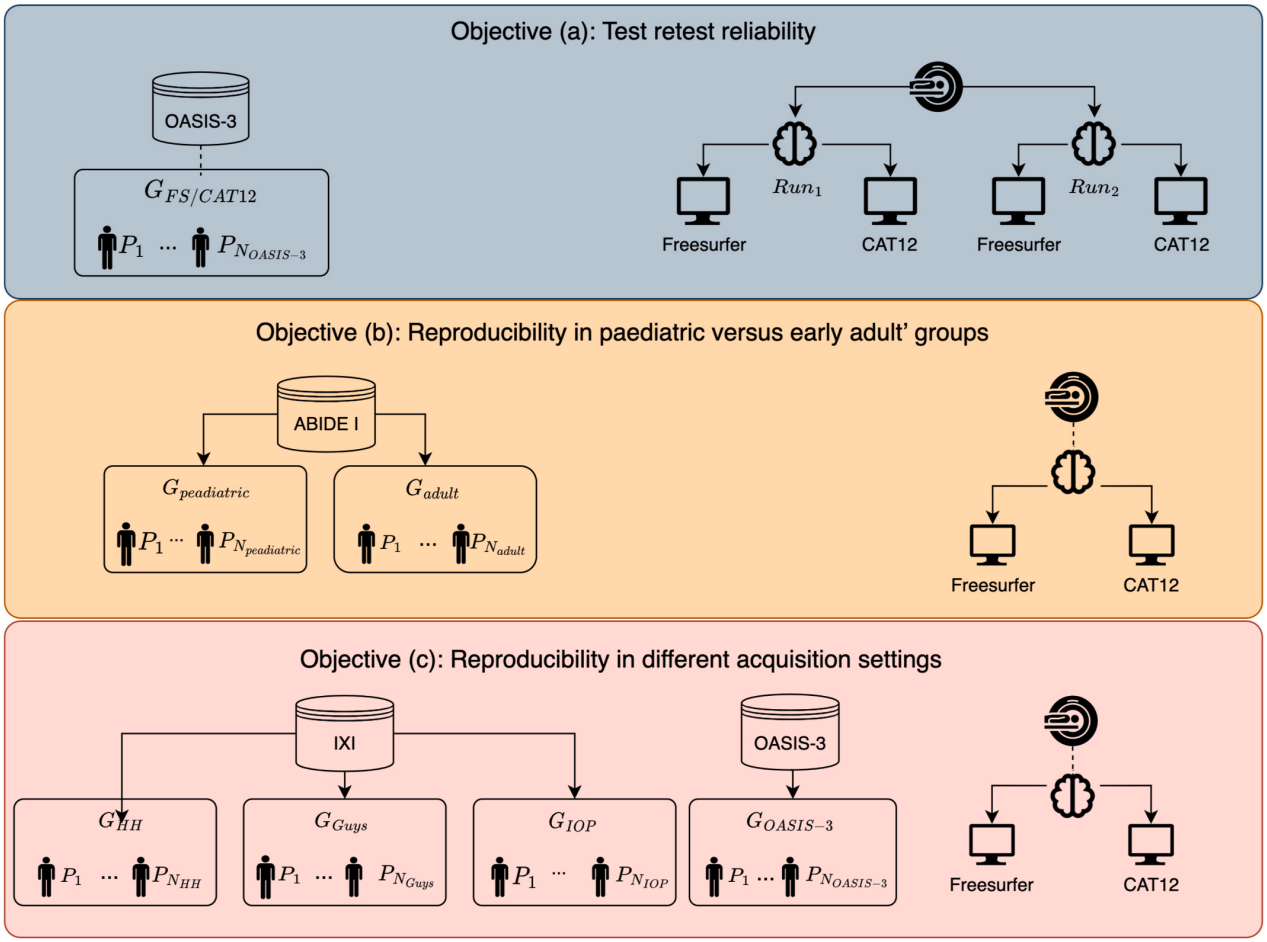


Fig. 1. Depiction of the objectives of this work as well as the data, groups (G) and pipeline used in each objective. In this figure N represents the total number of participants of a given group.

- Objective (c) assesses the reproducibility between FreeSurfer and CAT12 across different acquisition sites. In this objective two datasets were used: the IXI and OASIS-3 datasets, the former has three acquisition sites, whereas the latter has one acquisition site. Therefore, four groups were compared, one per acquisition site.

Fig. 1 summarizes the groups and data used for each objective.

2.1. Data

Three open sharing initiatives were used in this work: OASIS-3 (LaMontagne et al., 2019), IXI (https://brain-development.org/ixi-dataset/) and ABIDE I (Di Martino et al., 2014). The demographics statistics and the acquisition protocol are shown in Tables S1, S2 and S3 of the Supplementary material. Each of the open sharing initiatives used in this work is described below.

2.1.1. OASIS-3

OASIS-3 is an online repository that comprises images of the same individuals captured from multiple modalities, such as MRI and PET (Positron Emission Tomography). Each subject has one or more session, and a session may have multiple runs of the same MRI modality. All images of the OASIS-3 dataset were acquired in the same research centre over the course of 30 years, using different scanners. In this work, only images acquired with the Siemens TIM Trio 3T scanner were used. For the objective (c) all the subjects considered cognitively normal at the first MRI session were used, which performed a total of 494 participants, three participants had pre-processing errors, in total 491 subjects were considered. For the test-retest analysis it is required that each subject has two T1-weighted images acquired, at maximum, 60 min apart, 299 subjects fulfilled this requirement, three subjects were excluded from the subset due to pre-processing errors, thus in total 296 subjects were used in the analysis.

2.1.2. ABIDE I

ABIDE I is a multi-site data sharing initiative (17 sites in total) with neurotypical controls and subjects diagnosed with Autism Spectrum Disorder (ASD). In this study only the neurotypical controls were considered. Five subjects were discarded due to errors during the pre-processing step, in total 566 subjects were analysed.

2.1.3. IXI

IXI is a dataset collected from healthy adults in three distinct hospitals in London: the Hammersmith Hospital (HH), the Guy's Hospital (GH) and the Institute of Psychiatry (IOP). Philips MRI scanners with 3T and 1.5T were used to acquire brain images in HH and GH, respectively. The IOP images were acquired with a scanner from General Electrics (GE) with a field strength of 3T. In total, this dataset contains data collected from 562 subjects; from these a subset composed of 558 subjects were successfully pre-processed in this study, the demographics of the IXI dataset is in Table S2 of the Supplementary Material.

2.2. Frameworks

2.2.1. CAT12

CAT12 (http://www.neuro.uni-jena.de/cat/) is a computational framework to automatically pre-process T1-weigthed MR images. Its routines are implemented in MATLAB. In this work the default pre-processing routine (*Segment*) of CAT12 was used to pre-process the T1-wighted images. The CAT12, SPM12 and MATLAB version were version 1742, v7771 and R2020a (9.8.0.1323502), respectively. The backbone of the segmentation routine is the unified segmentation algorithm (Ashburner and Friston, 2005). The segmentation step relied on Tissue Probability Maps (TPM) to differentiate the brain tissues. The TPM must be properly adjusted for morphological changes that occur in the human brain during neurodevelopment (Wilke et al., 2008). CAT12 integrates the toolbox Template-O-Matic (TOM) to generate customized TPM based on the age and gender of the subjects in the dataset. A paediatric TPM was generated for the ABIDE I dataset, using all the individuals with age lower or equal to 18 years old. In the CAT12 pre-processing, in the paediatric group only, the customized TPM was then used.

2.2.2. FreeSurfer

FreeSurfer (Fischl, 2012) is an established computational framework in the neuroscience community (Yousaf et al., 2018). This framework contains a routine called *recon-all* which implements all the required steps to pre-process T1-weighted MR images and extract cortical thickness estimations. In this work the images where pre-processed using the recon-all routine of the FreeSurfer version 7.1.1.

2.2.3. MRIQC

The MRI quality control tool (MRIQC) (Esteban et al., 2017) is an open-source framework which extracts different quality metrics from the T1-weighted images. This framework was used to extract the quality metrics from the MRI images. The signal-to-noise ratio (SNR) is one of the metrics extracted and was the metric chosen to represent the quality of an image.

2.3. Cortical parcellation

A standard approach in cortical thickness analysis is the division of the cortical ribbon into several regions of interest (ROI) according to a given template. In this work, the cortical template considered was the Destrieux template (Destrieux et al., 2010). This template divides each hemisphere into 74 ROIs based on a sulco-gyral parcellation. The ROIs are then grouped into five lobes: frontal, insula, temporal and occipital, parietal and limbic. A ROI may be in more than one lobe, for instance the *anterior part of the cingulate gyrus and sulcus* is in the frontal and limbic lobe.

2.4. Reproducibility and reliability metrics

The reproducibility and reliability between and within CAT12 and FreeSurfer, respectively, is assessed in this work. The metrics extracted were the same for both: the coefficient of determination (R^2) of a regression and the intraclass correlation coefficient (ICC), more specifically the ICC(3,1) proposed by (Shrout and Fleiss, 1979). The computation methodology is explained in the following two sections.

2.4.1. Test-retest reliability

The test-retest reliability aimed to assess the reliability of cortical thickness estimations within framework over a short period of time (Run_1 and Run_2). A fitting was performed in which the Run_1 and Run_2 cortical thickness estimations were the dependent and independent variables, respectively, from the fitting the R^2 was extracted. The ICC was computed comparing the results of the two runs.

2.4.2. Reproducibility

To obtain reproducibility metrics between CAT12 and FreeSurfer a linear regression was fitted to the data in which the independent and dependent variables were the cortical thickness estimations extracted by FreeSurfer and CAT12, respectively. The R^2 of the fitting was one of the reproducibility metrics, the other was the ICC which was computed comparing the cortical thickness estimations extracted by both frameworks.

2.5. Statistics

Two analyses were carried out per objective: a "participant analysis" and a "ROI analysis". In the former reproducibility/reliability metrics were extracted per individual whereas in the second metrics were

computed per ROI.

For each objective an overall analysis, participant and ROI analysis was performed as described in subsections 2.5.1, 2.5.2 and 2.5.3, respectively.

A significance threshold of 0.05 was considered in all statistical analyses in this work.

2.5.1. Overall analysis

In this analysis all the cortical thickness estimations for all ROIs and individuals were considered. Two plots were used to show the relationship of two estimates of cortical thickness: a regression plot and a Bland-Altman plot. The former shows the relation of the two estimates, whereas the latter depicts the mean and mean difference of these estimations.

2.5.2. Participant analysis

The workflow of the participant analysis for the reliability/reproducibility study is depicted in Fig. 2. In this analysis the reproducibility/reliability metrics are computed per participant using all participant' ROIs. Therefore, each participant has two reproducibility/reliability metrics: R^2 and ICC. Then, two studies are carried out: a *metric analysis* and an *analysis of age effect on reproducibility/test-retest reliability metrics.*

In the metric analysis a statistical test was performed per reproducibility/reliability metric to compare the metric means of different groups. The statistical test performed depends upon the objective as follows:

- Objective (a) – to compare the test-retest reliability metric means of different frameworks a *paired t-test* was conducted;
- Objective (b) - to compare the reproducibility metric mean of paediatric and adults' groups an *ANCOVA* was performed, in which the dependent variable was the metric, whereas the factors were the age, SNR, the age-group and the acquisition setting;
- Objective (c) – to test the acquisition setting effect on the reproducibility metric an ANCOVA was performed in which the independent variables were the acquisition setting, age and SNR, whereas the dependent variable was the reproducibility metric in consideration.

The assessment of the age impact on the reproducibility/reliability metrics is designated in this work by *Analysis of age effect on reproducibility/test-retest reliability metrics*. The values from the metrics analysis were grouped by metric and group. Then another linear fit was performed in which the dependent variable was the metric whereas the independent variables were the age and SNR. An additional term, the acquisition setting, was used for the objective (b) which is the only objective that has multiple acquisition settings within a group (see Fig. 2). It was considered that age influenced the metric if the *p*-value of the second fit was lower than the significance value (*p-value*<0.05).

The relation of age and a metric is given by the age' slope (β_{age}) of the second fit. In this study, the β_{age} of the R^2 and *ICC* of the second fit is given by $\beta^{R^2}_{age}$, and β^{ICC}_{age}, respectively. To further verify whether the relationship between age and reproducibility/reliability metrics was the same in every group, a *independent samples t-test* corrected for multiple comparisons was performed, per metric, comparing the β_{age} of the different groups (see Fig. 2).

2.5.3. ROI analysis

This analysis aimed to understand the reproducibility/reliability across ROIs. To achieve this two studies were conducted: a *metric analysis* and *paired t-test for each ROI.*

In the metric analysis a linear regression was fitted for each ROI in

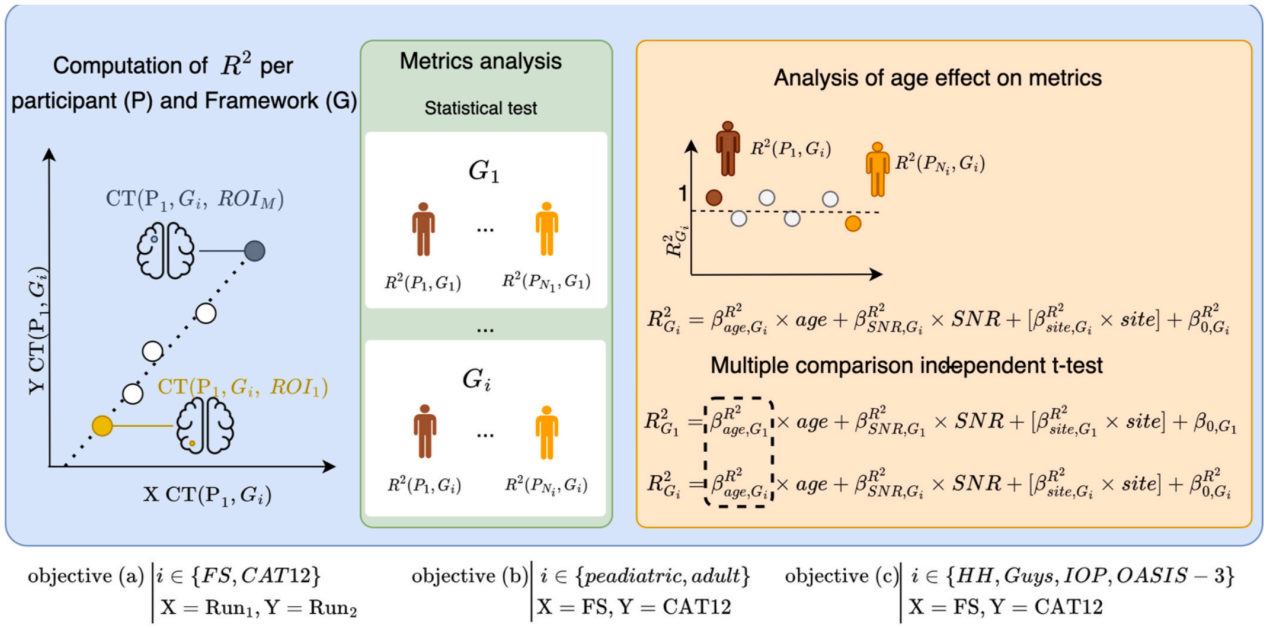


Fig. 2. Generic representation of the workflow of the Participant Analysis for the objectives proposed, considering the R^2 as the reliability/reproducibility metric. The first step of this analysis is the computation of the R^2 , per participant (P), using the cortical thickness (CT) estimations for that participant (blue block). In the objective (a) theR^2 is computed comparing the CT estimations of Run$_1$ and Run$_2$, whereas for the objective (b) and (c) the R^2 is computed using the CT estimations of FreeSurfer (FS) and CAT12. Then, a metrics analysis is performed: the R^2 means of different groups (G_i) are compared using the appropriate statistical test (green block). Finally, the analysis of age effect on metrics is done by performing a fit per group (orange block). The dependent variable is the R^2 and the independent variables are the age, SNR and site. It should be noted the site is only used in the objective (b) since it is the only one that has multiple sites within a group. The slopes of the age for the different groups, $\beta^{R^2}_{age,G_i}$, are compared using independent t-tests corrected for multiple comparisons. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

which the dependent and independent variables were the CAT12 and FreeSurfer estimations of cortical thickness for all individuals (see Fig. 3). To compare whether the reproducibility metrics across ROIs were identical in every group, the appropriate statistical test was performed:

- Objective (a) a paired-test was performed, per metric, to compare the ROI reliability metrics of the two frameworks.
- Objectives (b) and (c) a pairwise *independent samples t-test* was performed, and corrected with the Bonferroni method (Dubitzky et al., 2013).

Moreover, for the all objectives a *paired t-test* comparing the cortical thickness estimations of CAT12 and FreeSurfer was performed per ROI and group/framework.

2.5.4. Modelling age using cortical thickness

The reproducibility problem arises when different studies report different findings. The objective (c) aims to verify whether the use of different frameworks may have an impact in the same analysis. For this analysis, we considered the brain age problem (Franke and Gaser, 2019), in which a person's age is decoded from cortical thickness estimations, in our case. A multiple linear regression was fitted for each framework (CAT12 and FreeSurfer), in which the independent variables were the cortical thickness values for each ROI, and the dependent variable is the age. To test whether the reproducibility metric, in this case the mean reproducibility R^2, had an impact on the reported results we compared the ROIs which were considered significant in each framework model. The ROIs of each model were divided into overlapping and non-overlapping, the former are the ROIs which were significant in both framework models whereas the non-overlapping were the ROIs significant in only one framework model. An independent *t*-test was performed to verify whether the mean reproducibility R^2 was different in both groups (overlapping and non-overlapping).

3. Results

The results section is divided into 3 sub-sections, one for each objective:

- Objective (a) - Test-retest reliability
- Objective (b) - Reproducibility in paediatric versus early adult' groups
- Objective (c) - Reproducibility in different acquisition settings

Each subsection begins with an overview analysis of the data, then the individual analysis is presented, which is divided into the *metrics analysis* and *analysis of age and effect on the reproducibility/reliability metrics*. Finally, the results of the ROI analysis are described.

3.1. Test-retest reliability

In this section, the goal is to assess the test-retest reliability of each framework. For this, we used 296 subjects from the OASIS-3 dataset. The protocol followed in the OASIS-3 dataset implied that at least two MRI images per subject were acquired in the same day within a 60-min period. The two images per subject were pre-processed by both frameworks and the test-retest reliability within framework was assessed.

3.1.1. Overall analysis

The relationship between the cortical thickness extracted by the same framework in different images, independently of the framework is depicted in Fig. 4. The equation is given by $y = 0.95x + 0.11$ and $y = 0.96x + 0.08$, and the R^2 of the fitting is 0.92 and 0.93 for FreeSurfer and CAT12, respectively.

The results of the Bland-Altman plot show that the cortical thickness mean computed with each framework is not different across runs, being 2.38 and 2.36 for FreeSurfer and CAT12, respectively. Concerning the bias, the value is 0.005 mm and −0.0002 mm for FreeSurfer and CAT12, respectively. The former is significantly different from 0 ($t_{43807} = 9.39$,

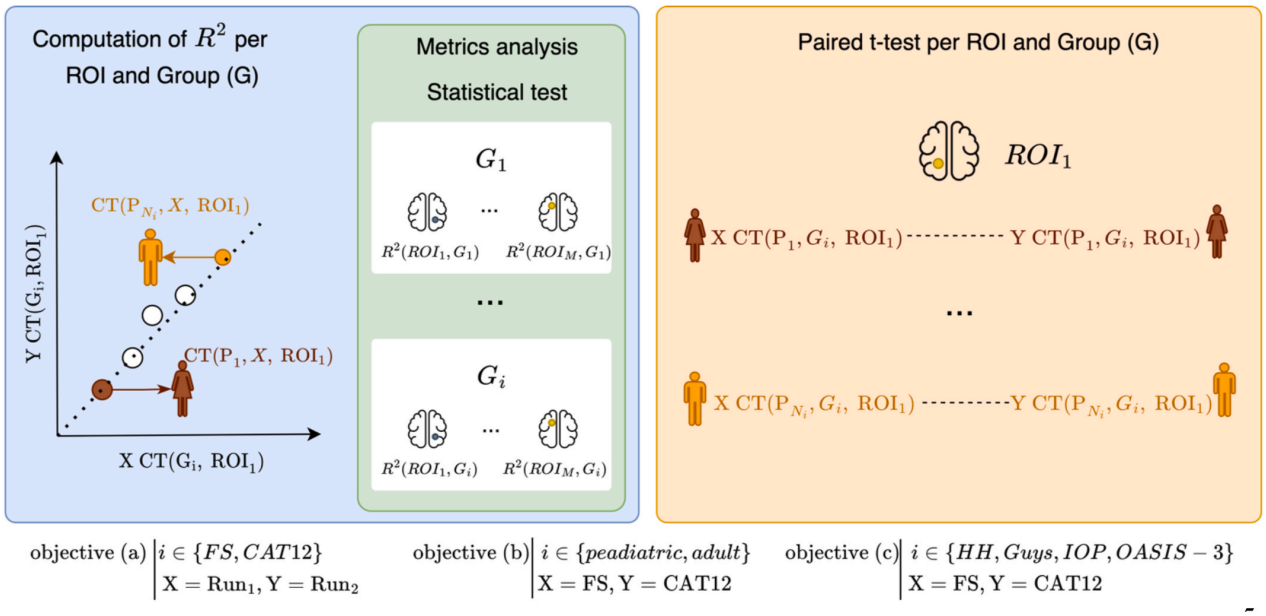


Fig. 3. Generic representation of the workflow of the ROI Analysis for the three objectives. In this figure the R^2 is used as the example for the reproducibility/reliability metric. A fit is performed using all the subject's cortical thickness estimations of a given ROI and from this fit the R^2 is extracted (blue block). The R^2 values of the different groups (G_i) are then compared using an appropriate statistical test (green block). Besides the fitting, a paired *t*-test (orange block) is performed, per ROI and group, comparing the cortical thickness estimations for different participants of the two runs, in the objective (a), and the two frameworks in the objectives (b) and (c). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

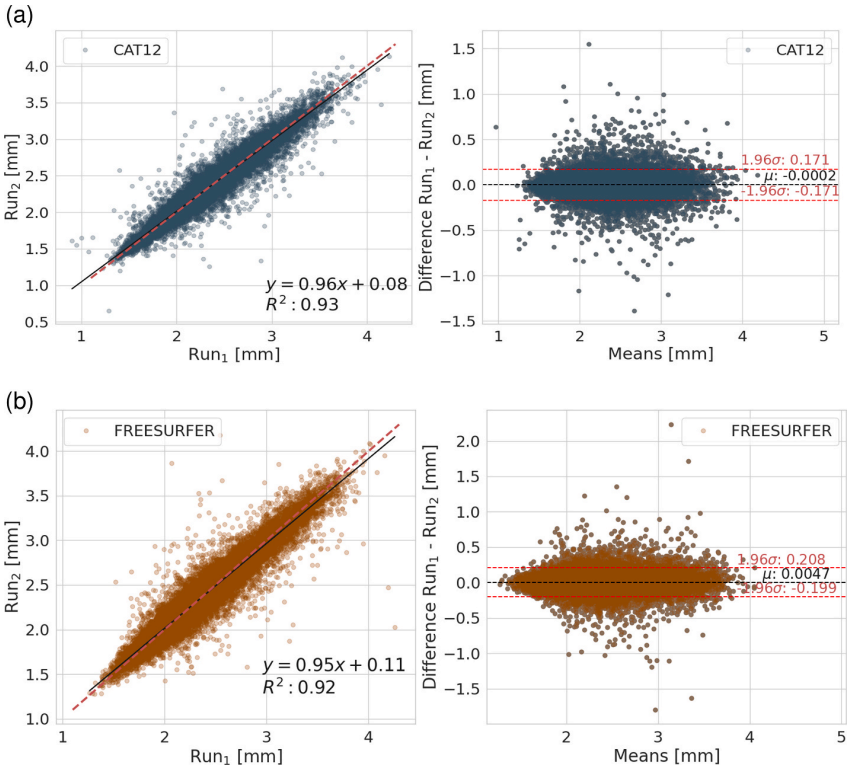


Fig. 4. Regression and Bland-Altman plot for (a) CAT12 and (b) FreeSurfer: On the left the distribution of cortical thickness extracted in Run 1 and Run 2, respectively, each point corresponds to a subject's ROI. The dashed red line on the left plot represents the equation y = x. On the right is the cortical thickness difference between runs. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Cohen'd:0.04) and the latter is not statistically significant. The confidence interval (CI) is [−0.20; 0.20] mm and [−0.17, 0.17] mm, therefore the CI width is 0.40 mm and 0.34 mm for FreeSurfer and CAT12, respectively.

3.1.2. Participant analysis

3.1.2.1. Test-retest metrics analysis. The reliability test-retest results for the participant analysis, per framework, are shown in Table 1.

The difference between the FreeSurfer and CAT12 test-retest reliability metrics' means of the frameworks is −0.01 for both R^2 and ICC. To verify whether the means are statistically different a *paired t-test* comparison was performed. The results suggest that the mean of FreeSurfer and mean of CAT12 is different for the R^2 ($t_{295} = 4.14, Cohen'd = 0.15$) and ICC ($t_{295} = 3.6, Cohen'd = 0.13$).

3.1.3. Analysis of age effect on test-retest metrics

The distribution of the test-retest reliability metrics with age is depicted in Fig. 5 and the statistical analysis results are shown in Table 2. The results show that age had no effect on CAT12 and FreeSurfer on test-retest reliability metrics.

Table 1
Test-retest metrics analysis, showing the mean and the standard deviation of R^2 and ICC, for the participant analysis, per framework.

Framework	R^2	*ICC*
FreeSurfer	0.93 ± 0.07	0.97 ± 0.04
CAT12	0.94 ± 0.06	0.96 ± 0.04

3.1.4. ROI level analysis

The reliability test-retest results, per framework, are presented in Table 3 and in Fig. 6, the detailed results of the R^2 grouped by lobe is in Table S4 whereas the R^2 and ICC by ROI are presented in Tables S5–S6 of the Supplementary Material. The test-retest reliability metrics are lower for the ROI analysis than for the participant analysis.

The comparison between the ROI test-retest reliability metric's mean values of FreeSurfer and CAT12 (*paired t-test*) showed a significant difference between the frameworks for R^2 (mean difference −0.08, $t_{147} = 9.83$, *Cohen'd*:0.83) and *ICC* (mean difference −0.06, $t_{147} = 9.28$, *Cohen'd*:0.81).

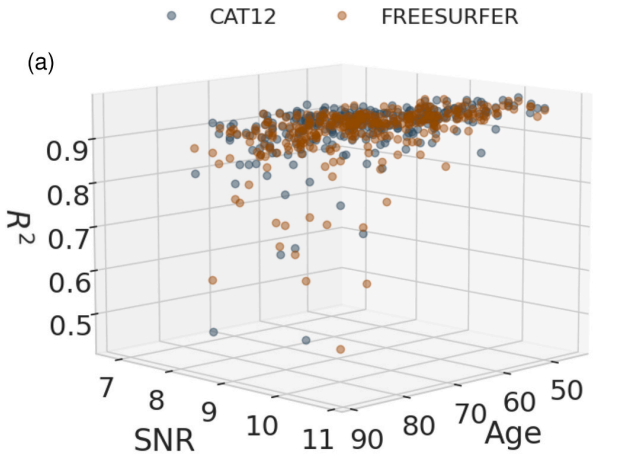
The individual ROI analysis revealed that the lobe with higher R^2 for CAT12 is the Parietal lobe (0.83) whereas for FreeSurfer is the temporal and occipital lobes (0.77). The lobe with lower R^2 is the frontal lobe for both CAT12 (0.76) and FreeSurfer (0.63).

The ROIs with higher difference between R^2 in test-retest reliability are the *parahippocampal gyrus, inferior occipital gyrus and sulcus, posterior-ventral part of the cingulate gyrus,* and the *planum polare of the superior temporal gyrus*.

Concerning the *paired t-tests performed for each ROI*, the results suggest that the cortical thickness estimations are similar between runs within each ROI, for all ROIs except the *Left Short insular gyri* in FreeSurfer, detailed results are in Table S7 of the Supplementary Material.

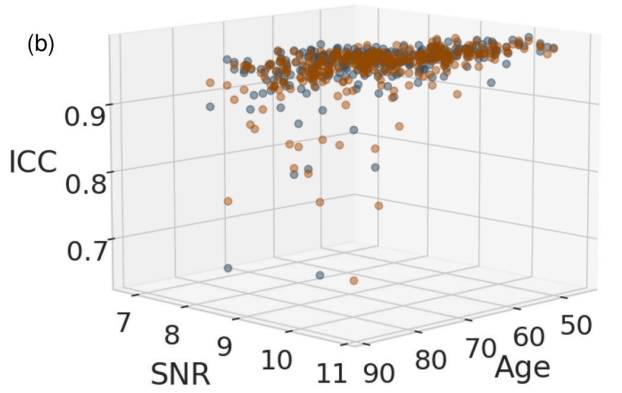
3.2. Reproducibility in paediatric versus early adults' groups

This section investigates whether different age stages (paediatric versus adults) have different reproducibility behaviour. The dataset used was the ABIDE I dataset, and the age range of the participants is from 6.5 to 56.2 years, and the analysis was carried out for two age groups: paediatric and adults.



$$R^2_{CAT12} = 6.4 \times 10^{-4} age + 4.6 \times 10^{-2} SNR + 4.6 \times 10^{-1}$$

$$R^2_{FREESURFER} = 3.3 \times 10^{-4} age + 4.7 \times 10^{-2} SNR + 4.6 \times 10^{-1}$$



$$ICC_{CAT12} = 3.9 \times 10^{-4} age + 2.6 \times 10^{-2} SNR + 7.0 \times 10^{-1}$$

$$ICC_{FREESURFER} = 2.1 \times 10^{-4} age + 2.6 \times 10^{-2} SNR + 7.0 \times 10^{-1}$$

Fig. 5. Distribution of the R^2 (a) and ICC (b) with age for each of the frameworks. Each point represents the (a) R^2 or (b) ICC value, at the participant level, of the cortical thickness test-retest reliability of each framework.

Table 2
Analysis of age effect on test-retest metrics: results of the regression with each of the reproducibility metrics: R^2($\beta^{R^2}_{age}$) and ICC (β^{ICC}_{age}).

	Framework	$\beta^{R^2}_{age}$	β^{ICC}_{age}
Age	*FreeSurfer*	3.2×10^{-4} ($p = 0.485$)	2.2×10^{-4} ($p = 0.445$)
	CAT12	6.4×10^{-4} ($p = 0.1$)	3.9×10^{-4} ($p = 0.121$)

Table 3
Mean and standard deviation of the $R^2$and ICC values for the test-retest reliability metrics ROI analysis.

Framework	R^2	*ICC*
FreeSurfer	0.70 ± 0.10	0.84 ± 0.06
CAT12	0.78 ± 0.09	0.88 ± 0.05

3.2.1. Overall analysis

All the cortical thickness values extracted by both frameworks are depicted in Fig. 7. The equation that translates the relationship between

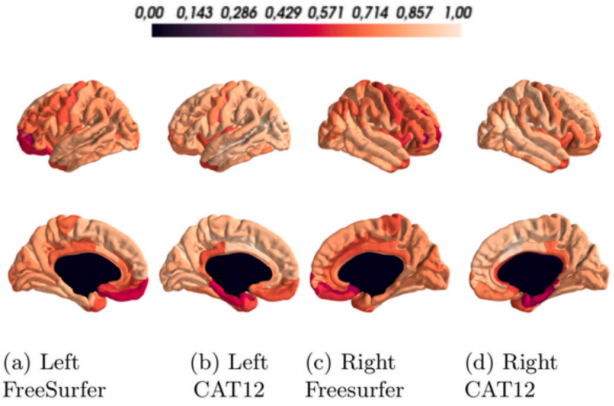


Fig. 6. R^2 values distribution between the cortical thickness extracted by FreeSurfer or CAT12 values per ROI, per hemisphere.

FreeSurfer and CAT12 values is $CAT12 = 0.61 \times FREESURFER + 0.92$ and $CAT12 = 0.69 \times FREESURFER + 0.75$, for paediatric and adult samples, respectively, and the R^2 is 0.48 and 0.64. The mean difference between the estimations is 0.065 ± 0.28 mm. The Bland-Altman plot shows a FreeSurfer (CAT12) mean of cortical thickness of 2.60 (2.52) mm and 2.49 (2.45) mm for paediatric and adult samples, respectively. Concerning the cortical thickness difference, the CI is between [−0.50, 0.66] mm and [−0.44,0.51] mm, thus a CI width of 1.16 mm and 0.95 mm for paediatric and adult samples. The bias mean is 0.080 mm and 0.034 mm, for paediatric and adult samples, respectively. The *t*-test suggests that the bias is statistically different from zero in both cases (paediatric, $t_{56831} = 6.47, cohen'd: 0.27$ and adult, $t_{26935} = 2.3$, $Cohen'd: 0.14$).

3.2.2. Participant analysis

3.2.2.1. Reproducibility metrics analysis. The results observed when analysing the two age groups are summarized in Table 4. The difference between the paediatric and adult group is −0.13 and −0.10 for R^2 and *ICC*, respectively. Three ANCOVAs were performed, one per reproducibility metric. The ANCOVA results are shown in Tables S8 and S9 in the Supplementary Material. In summary, in the two metrics all the independent factors (age, SNR, age-group and acquisition setting) are suggested to influence the reproducibility metrics. The SNR was the factor with higher dispersion, followed by age, acquisition setting and age-group.

3.2.2.2. Analysis of age effect on reproducibility metrics. The evolution of reproducibility metrics with age and SNR per acquisition setting, is depicted in Fig. 8. It shows an interesting and unexpected pattern: in children and adolescents all reproducibility metrics improve with age. Moreover, the variability decreases with age as well as the worst attained value, for each reproducibility metric.

Table 5 shows the relation between the age for each of the reproducibility metrics. In the paediatric group the relation was found to be significant, for both metrics (R^2 and ICC) and age. Regarding the β_{age}comparisons, the *independent t-test* suggested that the difference between paediatric and adult group ($\beta_{age,R^2}: 9.1 \times 10^{-3}$ and $\beta_{age,ICC}$: 6.0×10^{-3}) of relation between age and reproducibility metrics is different between the two age groups (paediatric vs adult) for the two reproducibility metrics: $slope_{R^2}(t_{564} = 25.96$, *Cohen'd: 2.34*) and $slope_{ICC}(t_{564} = 23.72, Cohen'd: 0.69)$.

3.2.3. ROI level analysis

The reproducibility metrics for the ROI analysis are depicted in

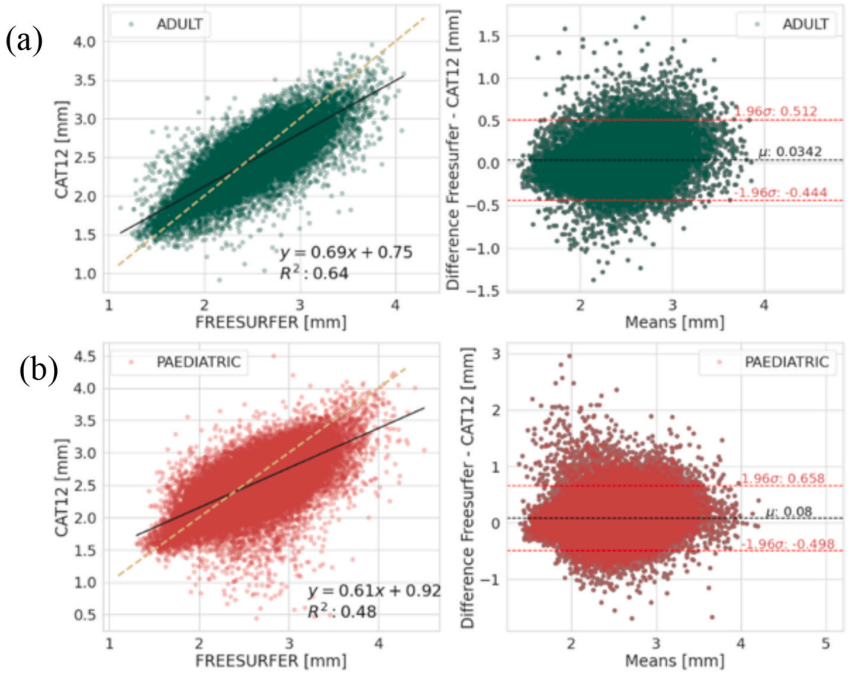


Fig. 7. Regression and Bland-Altman plot for the paediatric (a) and adults' (b) analysis. On the left the distribution of cortical thickness extracted with FreeSurfer and CAT12, respectively, each point corresponds to a subject's ROI. The dashed yellow line on the left plot represents the equation y = x. On the right the cortical thickness means and difference (FreeSurfer – CAT12) of the frameworks. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Reproducibility metrics analysis: Mean+/-standard deviation of R^2 and ICC, for the individual analysis, grouped by age (greater or less than 18 years) for the neurodevelopment analysis.

Age (years)	R^2	ICC
≤ 18	0.51 ± 0.14	0.69 ± 0.11
> 18	0.64 ± 0.09	0.79 ± 0.06

Fig. 9. Table 6 presents a summary of the metrics, per age group (the detailed lobes results is in Table S10 and the results per ROI for R^2 and ICC is in Tables S11 and S12, respectively, of the Supplementary Material).

To verify whether the difference between the paediatric and adult age groups was significant for the R^2 (-0.08) and $ICC(-0.07)$ an *independent samples t-test* was performed. The results showed differences between age groups, for the two metrics: R^2 $(t_{294} = -4.79, Cohen'd: 0.56)$ and ICC $(t_{294} = -3.98, Cohen'd: 0.46)$.

The lobe with higher difference between the age groups is the temporal and occipital lobe whereas with the lowest difference is the frontal lobe. The parietal lobe is the one with higher mean R^2, 0.49 and 0.41 for adults and paediatric, respectively. Whereas insula lobe is the one with lower mean R^2, 0.24 and 0.16 for adults and paediatric, respectively. The ROI with higher reproducibility in children and adolescents is the *postcentral sulcus*, which is also the one of the ROIs with lowest difference between the age groups.

The ROI *paired t-tests* showed that 137 ROIs had at least one age-group with differences in the cortical thickness estimations of both frameworks. In the children and adult group, 130 and 109 ROIs had differences between the cortical thickness estimations extracted by Freesurfer and CAT12, respectively. The *p*-values were corrected for the 296 comparisons, detailed results in Table S13 of the Supplementary Material.

3.3. Reproducibility in different acquisition settings

This section aims to understand whether the acquisition setting and age has impact on reproducibility metrics on mature brains. In this analysis two open-data sharing initiatives were used: IXI and OASIS-3. The analysis was conducted per acquisition setting. Therefore, four groups were analysed: three groups for the IXI dataset (one per site/acquisition setting) and one group for the OASIS-3. Furthermore, the reproducibility between CAT12 and FreeSurfer metrics was also computed for a subset of OASIS-3. The subset was the one used in test-retest reliability analysis, objective (a), which includes subjects who have two T_1-weighted images. In this case the cortical thickness was averaged over the two runs and the reproducibility metrics were computed.

3.3.1. Overall analysis

In this case, given that there are four acquisitions settings we decided to report the Bland-Altman results independently of the acquisition setting, the detailed results per acquisition setting are in Fig. S1 of the Supplementary Material. Considering all the adult individual ROIs, the overall reproducibility metrics were extracted, which are shown in Fig. 10 with a Bland-Altman plot. The linear regression has the following equation: $CAT12 = 0.70 \times FREESURFER + 0.68$, with a R^2 of 0.65. Regarding the Bland-Altman analysis the mean cortical thickness value is 2.39 mm and 2.36 mm for FreeSurfer and CAT12, respectively. Regarding the cortical thickness difference between the frameworks the mean difference is 0.032 which is statistically different from zero $(t_{155251} = 5.30, Cohen'd: 0.13)$. Regarding the CI the values range from [−0.44; 0.50] mm which is a width of 0.93 mm.

3.3.2. Participant analysis

3.3.2.1. Reproducibility metrics analysis. The overall R^2 for the adult datasets is 0.65, for the participant level analysis, and Table 7 summarizes the reproducibility metrics per acquisition setting. The acquisition setting with higher reproducibility metrics is the Guys, followed by OASIS-3, HH and IOP.

The ANCOVA suggested an acquisition setting and SNR effect in both metrics, yet the acquisition setting has the highest factor of dispersion dependent of the metric. The age had no effect in both metrics. The post-

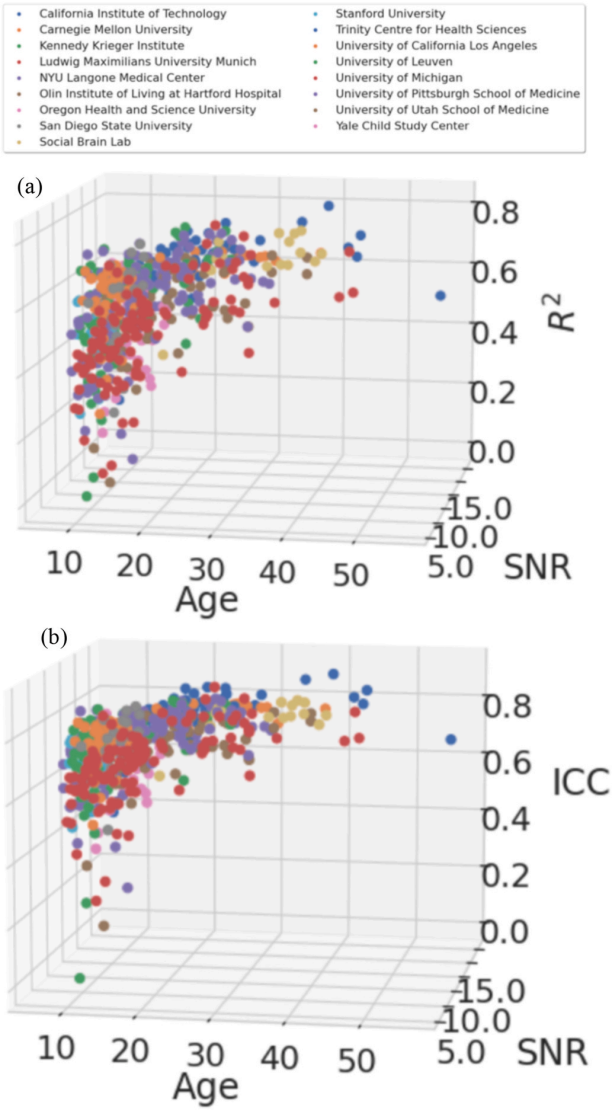


Fig. 8. R^2 and ICC by age and SNR for each site of the ABIDE I.

Table 5
Reproducibility metrics and age relation analysis: regression between age and SNR and each of the reproducibility metrics R^2($\beta_{age}^{R^2}$) and ICC (β_{age}^{ICC})per age group. The significant values are shown in bold.

	Age (years)	$\beta_{age}^{R^2}$	β_{age}^{ICC}
Age	≤ 18	$\mathbf{1.2 \times 10^{-2}}$ ($p = 8.6 \times 10^{-7}$)	$\mathbf{8.9 \times 10^{-3}}$ ($p = 1.3 \times 10^{-5}$)
	> 18	$\mathbf{2.9 \times 10^{-3}}$ ($p = 1.2 \times 10^{-3}$)	$\mathbf{2.9 \times 10^{-3}}$ ($p = 1.2 \times 10^{-3}$)

hoc comparisons revealed differences between all acquisition settings, detailed results are in Tables S14–S17 in the Supplementary Material. In Fig. 11 the acquisition setting effect is quite visible, i.e., the IOP reproducibility metrics are clustered apart from the Guys, HH and OASIS-3.

The reproducibility results for the OASIS-3 test-retest subset were 0.67 ± 0.07 and 0.81 ± 0.06 for the R^2 and ICC, respectively.

3.3.2.2. Analysis of age effect on reproducibility metrics. Fig. 11 and

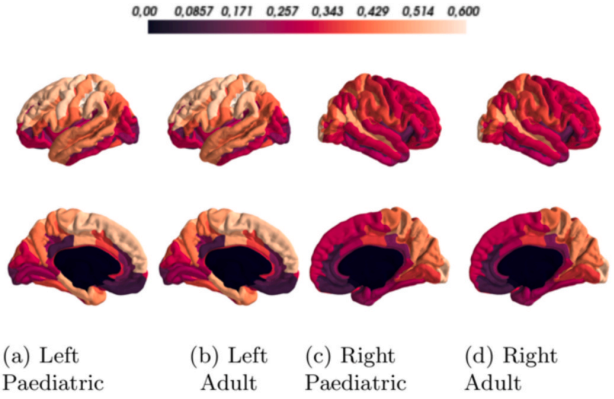


Fig. 9. R^2values distribution between the estimations of cortical thickness extracted by Freesurfer and CAT12 values per ROI, hemisphere, and per age group.

Table 6
Mean and standard deviation of R^2 *and ICC*, for the ROI analysis, grouped by age (greater or less than 18 years) for the neurodevelopment analysis.

Age (years)	R^2	ICC
≤ 18	0.30 ± 0.14	0.52 ± 0.14
> 18	0.38 ± 0.16	0.59 ± 0.15

Table 8 summarize the assessment of the impact of age and acquisition setting on reproducibility metrics. The results are consistent independently of the metric. The statistical analysis revealed an age effect on Guys and OASIS3 sites.

The comparative analysis of the β_{age} ($\beta_{age}^{R^2}$; β_{age}^{ICC}) of each acquisition setting points out that the two $\beta_{age}$are indeed different between all acquisition settings, except between HH and IOP, detailed results are in Tables S18–S19 of the Supplementary Material. Therefore, the relation between age and the reproducibility metrics is dependent on the acquisition setting.

3.3.3. ROI level analysis

A summary of the reproducibility metrics results for the ROI analysis, per acquisition setting, is shown in Table 9. The results show that the HH, Guys and OASIS-3 have better reproducibility metrics than the IOP.

Regarding the reproducibility results for ROI analysis and using the OASIS-3 test-retest dataset the R^2 and ICC were 0.40 ± 0.21 and 0.60 ± 0.21, respectively.

The R^2 values per ROI for each open sharing initiative are depicted in Fig. 12 and the ROI and lobe R^2 values are in Tables S20–S21 and the ICC values per ROI are in Tables S22–S23 of the Supplementary Material. The R^2 differs throughout the brain, it is clear that there are regions with higher reproducibility metrics than others. Yet, the ROI reproducibility pattern is similar across acquisition settings. In general, while the parietal lobe has the higher mean R^2, the limbic lobe has the lower mean R^2. The subcallosal area and anterior transverse collateral sulcus (both ROIs are near the corpus callosum) are the ROIs with the lowest R^2.

Most ROIs have a similar R^2 value for the right and left hemisphere, yet there are discrepancies. The *transverse temporal sulcus*, *middle occipital sulcus, lunatus sulcus* and *orbital gyri* yield a difference in R^2 between hemispheres higher than 0.5, independently of the acquisition setting. Nevertheless, in some ROIs, such as the *middle frontal gyrus*, the difference between hemispheres is only observed in the OASIS-3 dataset but not in others.

Four *paired t-test* were performed per ROI, one per acquisition setting, the *p-values* were corrected for the 576 comparisons. The results

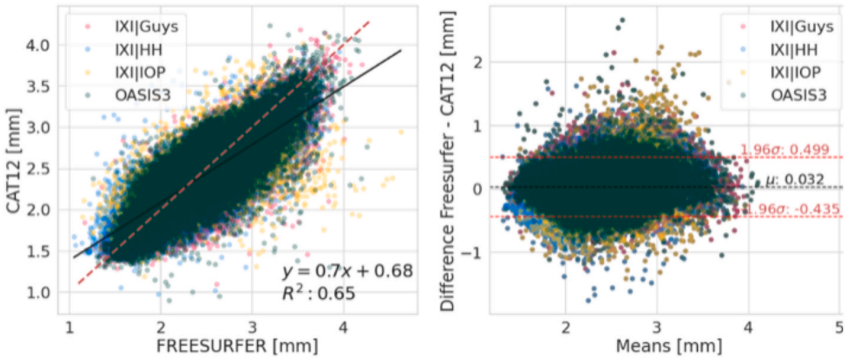


Fig. 10. Regression and Bland-Altman for the adult dataset with different acquisition settings. On the left the distribution of cortical thickness extracted with FreeSurfer and CAT12. The dashed red line on the left plot represents the equation y = x. On the right is the plot with the mean and difference (FreeSurfer – CAT12) between the cortical thickness estimates of the frameworks. In this plot each point corresponds to a participant's ROI and different the acquisition settings are coloured differently. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 7
Reproducibility metrics analysis: mean and standard deviation of R^2 and ICC for the participant analysis, grouped by acquisition setting for the IXI and OASIS-3 datasets.

Acquisition setting	R^2	ICC
Guys	0.69±0.06	0.82 ± 0.04
HH	0.65±0.09	0.79 ± 0.07
IOP	0.38±0.10	0.60 ± 0.08
OASIS-3	0.65±0.08	0.80 ± 0.06

revealed differences between the frameworks on 145 ROIs, in at least one acquisition setting. Briefly, 132, 116, 94 and 130 ROIs were significantly different in Guys, HH, IOP and OASIS-3, respectively. Detailed results are presented in Table S24 of the Supplementary Material.

3.3.3.1. Modelling age using cortical thickness values. Two models were trained, one per framework, with adult data from the IXI and OASIS-3 datasets. The adjusted R^2 for FreeSurfer and CAT12 models is 0.71 and 0.73, respectively. The number of ROIs considered significant for age prediction is 29 and 38 for FreeSurfer and CAT12, respectively, the details are in Tables S25 and S26 in the Supplementary Material. An analysis was carried out to verify which ROIs were significant in age prediction in both frameworks. The results point out that there are 9 overlapping ROIs, regions are significant in age prediction using both frameworks for estimation of cortical thickness. The overlapping ROIs and their mean reproducibility, R^2, across sites (which is represented between the parentheses) are: *left superior frontal gyrus (0.66), left triangular part of the inferior frontal gyrus (0.60), right angular gyrus (0.57), right superior occipital sulcus and transverse occipital sulcus (0.48), right superior part of the precentral sulcus (0.43), right occipital pole (0.34), left lingual gyrus (0.27), right anterior transverse temporal gyrus (0.28)* and the *left medial orbital sulcus (0.03)*.

In Fig. 13 is depicted the distribution of the reproducibility metric values, R^2, for each framework for the two groups: overlapping and non-overlapping ROIs. The independent *t*-test performed, per framework, revealed no difference between the groups for neither framework.

4. Discussion

This study reveals that the test-rest reliability of cortical thickness estimations within two computation frameworks (CAT12 and FreeSurfer) attains higher values than the reproducibility between frameworks, as expected. Concerning the test-retest reliability the Bland-Altman analysis suggested that the bias is only significant in the FreeSurfer framework, yet the effect is small. Furthermore, the dispersion for CAT12 is lower than for FreeSurfer. The cortical thickness estimation mean was identical between frameworks (2.36 mm and 2.38 mm), these mean values are close to previous report mean cortical thickness (Hutton et al., 2008). The confidence interval was 0.34 mm and 0.40 mm, respectively, which corresponds to an error of 14% and 17% for a mean cortical thickness of 2.37 mm.

At the individual level, the CAT12 attains statistically higher test-retest metrics than FreeSurfer, despite a very small 0.01 difference, which may be negligible from the biological point of view. An age effect was not detected in neither CAT12 nor FreeSurfer test-retest reliability metrics. Nevertheless, a SNR effect was detected in the two test-retest reliability metrics, therefore the performance within the framework is sensitive to the image quality. The positive slope between the SNR and the metrics reveals that images with higher quality have higher reliability. An analysis of the relation between age and SNR showed that older individuals have lower SNR values. This is in agreement with the literature reporting that motion artifacts are higher in older individuals and, which in turn affect cortical thickness estimation (Madan, 2018). At the ROI level the test-retest difference between frameworks is higher (0.08 for R^2) and was found to be statistically significantly for all measures with a strong effect on ICC and R^2. Once again, the CAT12 has higher test-retest metrics than FreeSurfer. Regarding test-retest reliability per lobes the parietal lobe and the temporal and occipital lobes have higher test-retest reliability. In contrast the frontal lobe has lower reproducibility in both frameworks.

This study reveals a significant discrepancy between the cortical thickness reproducibility metrics extracted by FreeSurfer and CAT12 in adults and paediatric populations. The reproducibility metrics are significantly lower in children and adolescents' than in adults, which may impact on the interpretation of neurodevelopmental studies. Furthermore, the reproducibility metrics increase with age until the age of 18 years old and stabilize in adulthood (Fig. 8). The reason behind this result might be nonlinear brain maturation, which changes brain morphology in children/adolescents (Gibb and Kovalchuk, 2018). These changes are a source of variability that might increase the difficulty to accurately estimate cortical thickness. Besides maturation, another cause for the low reproducibility in paediatric group might be motion, which tend to be higher in younger individuals (Pardoe et al., 2016) and causes artifacts that affect the accurate extraction of cortical thickness (Gilmore et al., 2021). It should be noted that age had an effect in children's cortical thickness reproducibility even when our models introduce SNR as a covariate. Thus, we believe some of the reproducibility issues might be inherent to biological variability introduced by the brain maturation process. These results are pursuant with published findings where it is shown that different pre-processing pipelines, in early childhood, yield different results (Haynes et al., 2020). The low reproducibility between pipelines might be an explanation for cortical thickness inconsistencies found in the neurodevelopment phase in health and disease (Bedford et al., 2020; Khundrakpam et al., 2017; Van Rooij et al., 2018; Walhovd et al., 2016; Wallace et al., 2010).

Concerning the ROI analysis, the overall pattern of reproducibility

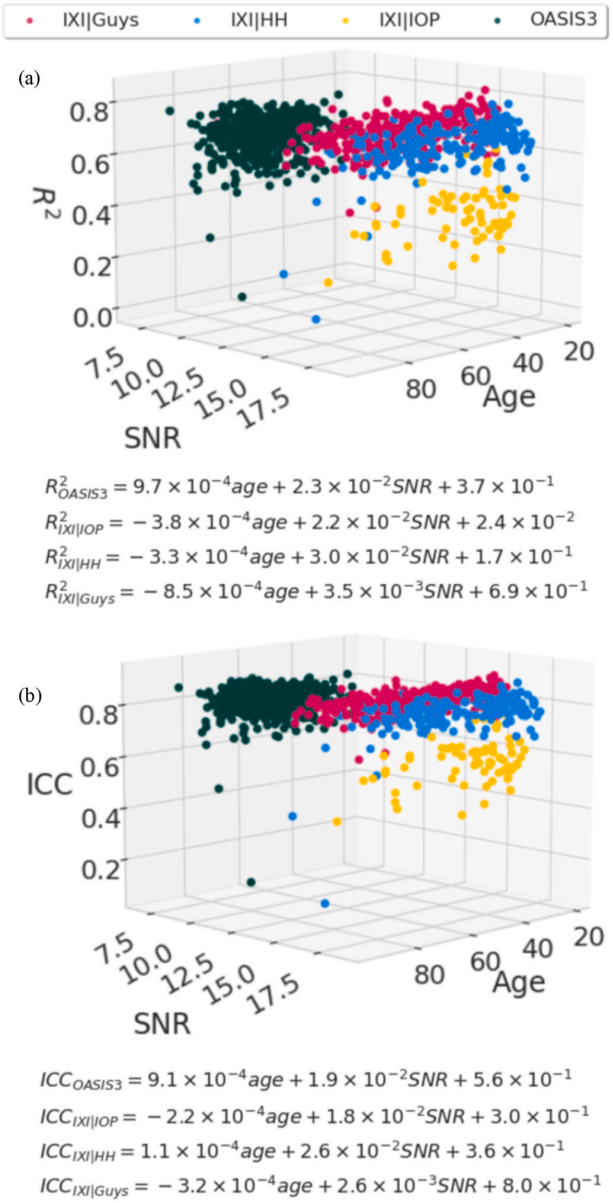


Fig. 11. Reproducibility metrics (R^2 and ICC) at the participant level and its relationship with age and SNR for the different sites of the IXI and the OASIS-3 datasets.

metrics is the same in both age groups: the parietal lobe is the region with higher reproducibility whereas the insula is the one with lower reproducibility. There is previous evidence that the insula has high annual changes in cortical thickness during childhood and adolescence whereas in parietal cortex the annual changes are less exacerbated (Sowell et al., 2004). Therefore, the reproducibility might be related to the annual change rate of cortical thickness, the higher is the rate of annual changes the lower is the reproducibility.

In adults, the reproducibility between frameworks slowly decreases with age, i.e., older adult individuals exhibit lower reproducibility metrics than younger adult individuals, in some acquisition settings. An age effect was detected, for both reproducibility metrics, in Guys and OASIS3 acquisition settings, even though we controlled for the image SNR. The reproducibility decrease with age may be related to the fact

Table 8
Reproducibility metrics and age relation analysis of the regression of age with each of the reproducibility metrics R^2 ($\beta^{R^2}_{age}$) and ICC (β^{ICC}_{age}), per acquisition setting. The significant values are shown in bold.

	Acquisition setting	$\beta^{R^2}_{age}$	β^{ICC}_{age}
Age	Guys	$\mathbf{-8.5 \times 10^{-4}}$ ($p = 1.3 \times 10^{-4}$)	$\mathbf{-3.2 \times 10^{-4}}$ ($p = 2.3 \times 10^{-2}$)
	HH	-3.3×10^{-4} ($p = 4.2 \times 10^{-1}$)	1.1×10^{-4} ($p = 7.5 \times 10^{-1}$)
	IOP	-3.8×10^{-4} ($p = 6.1 \times 10^{-1}$)	-2.2×10^{-4} ($p = 7.1 \times 10^{-1}$)
	OASIS-3	$\mathbf{9.7 \times 10^{-4}}$ ($p = 1.8 \times 10^{-2}$)	$\mathbf{9.1 \times 10^{-4}}$ ($p = 2.9 \times 10^{-3}$)

Table 9
Mean and standard deviation of R^2 *and ICC*, for the ROI analysis, per acquisition setting for the IXI and OASIS-3 datasets.

Site	R^2	*ICC*
Guys	0.40 ± 0.19	0.60 ± 0.18
HH	0.39 ± 0.19	0.59 ± 0.18
IOP	0.24 ± 0.16	0.44 ± 0.18
OASIS-3	0.39 ± 0.21	0.59 ± 0.20

that older individuals move more in the scanner than younger individuals, which adds noise to the images and, consequently, the methods have more difficulties in extracting accurate metrics (Pardoe et al., 2016; Reuter et al., 2015). Additionally, the changes in brain morphometry that occur naturally with the ageing process might also underly the decrease in reproducibility of these measures with increasing age. The lack of an age effect on the IOP might be related to its sample size (the IOP has 67 participants, whereas the HH, Guys and OASIS-3 have 179, 312 and 491 participants, respectively).

Besides the age effect, an acquisition setting effect was detected on the reproducibility metrics, which translates that sites with a similar SNR distribution had different reproducibility distributions. Furthermore, the relationship between age and reproducibility metrics is dependent on the acquisition setting. Therefore, the low reproducibility may be related to the acquisition setting that affected all the images in the same way, yet each acquisition setting affects the images differently, and the frameworks were unable to overcome this limitation. Moreover, the test-retest results also suggest that the reproducibility metrics are similar even when the cortical thickness estimation is averaged across two images of the same subject. This was as expected given the high values of the test-retest reliability of each frameworks. The current results highlight the challenge of multicentric studies. The conclusions of these large (typically consortia) studies might be vulnerable to acquisition settings imprints on the data and must be interpreted with caution. Nonetheless, multi-centre studies are crucial to hypothesis generalization. Therefore, data harmonization techniques should ideally be available in pre-processing frameworks to reduce the acquisition setting impact on the images.

The reproducibility at the individual level is higher than at the ROI level. This is expected since at the individual level there is only variance related to the framework whereas at the ROI level analysis there is variance related to both the participant and framework. This finding highlights the fact that the frameworks are not capable to overcome the individual variability and cross-sectional studies are more susceptible to inconsistencies leading to inaccurate conclusions. Nevertheless, the reproducibility at the ROI level is consistent across the four acquisition settings. This finding corroborates the results published in (Masouleh et al., 2020) in which it is shown that the ROI reproducibility is similar across datasets.

The analysis of the outcome of the age prediction problem showed

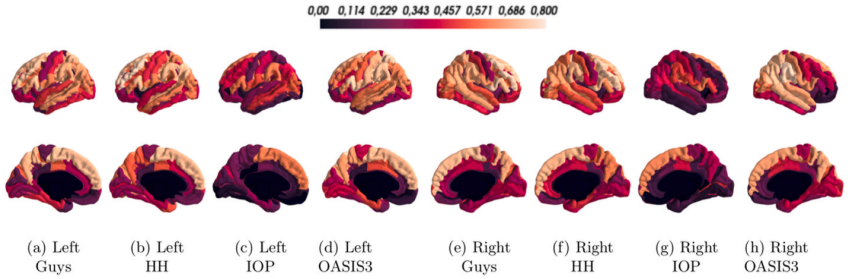


Fig. 12. R2 values distribution between the cortical thickness extracted by FreeSurfer and CAT12 values per ROI, hemisphere and acquisition setting.

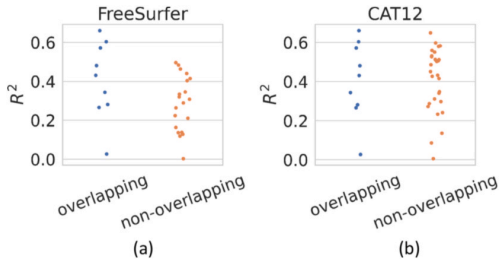


Fig. 13. Distribution of the mean reproducibility, R^2, of the overlapping and non-overlapping ROIs for FreeSurfer (a) and CAT12 (b).

that CAT12 and FreeSurfer yield different results. Nonetheless, despite the differences, the results also show that some ROIs with high reproducibility between FreeSurfer and CAT12 overlap across frameworks. Yet, a high reproducibility value does not guarantee the same result in both models, as there are ROIs with high values of reproducibility across frameworks that are only significant in predicting age in one of the frameworks.

5. Conclusion

This work shows that the reproducibility of cortical thickness estimations with CAT12 and FreeSurfer is weaker at the ROI-level than at the subject-level. These findings show that the inference from cross-sectional studies, in which conclusions at the ROI level are made from distinct sites, groups and frameworks, might explain inconsistencies reported in the literature. Furthermore, conclusions from metanalyses in which the pre-processing protocol was not the same might be compromised. Thus, the pre-processing framework should be considered when different studies are being compared. Our results show that the test-retest reliability is reassuringly high at the individual level for images obtained using the same acquisition protocol. Consequently, the conclusions from longitudinal studies, with the same acquisition and pre-processing protocol are less sensitive to the described processing framework effects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this work is from open data repositories. The code is available at: https://cibit-uc.github.io/fs-cat12-cortical-thickness-reproducibility/intro.html

Acknowledgements

This work was supported by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit (UIDB/00326/2020 or project code UIDP/00326/2020) and within the scope of CIBIT R&D unit (FCT, grant numbers DSAIPA/DS/0041/2020 and UIDB/4950/2020 and UIDP/4950/2020). FCT also funded a doctoral grant to Maria de Fátima Dias (SFRH/BD/145798/2019) and an individual contract to João Valente Duarte (Individual Scientific Employment Stimulus 2017 - CEECIND/00581/2017). Furthermore, we are thankful to the Laboratório de Computação Avançada, Universidade de Coimbra for granting us access to the supercomputer on which the data were pre-processed.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ynirp.2022.100137.

References

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018.

Bedford, S.A., Park, M.T.M., Devenyi, G.A., Tullo, S., Germann, J., Patel, R., Anagnostou, E., Baron-Cohen, S., Bullmore, E.T., Chura, L.R., Craig, M.C., Ecker, C., Floris, D.L., Holt, R.J., Lenroot, R., Lerch, J.P., Lombardo, M.V., Murphy, D.G.M., Raznahan, A., Ruigrok, A.N.V., Smith, E., Spencer, M.D., Suckling, J., Taylor, M.J., Thurm, A., Lai, M.C., Chakravarty, M.M., 2020. Greater cortical thickness in individuals with ASD. Mol. Psychiatr. 2020 (253 25), 507–508. https://doi.org/10.1038/s41380-020-0691-y.

Burton, E.J., McKeith, I.G., Burn, D.J., Williams, E.D., O'Brien, J.T., 2004. Cerebral atrophy in Parkinson's disease with and without dementia: a comparison with Alzheimer's disease, dementia with Lewy bodies and controls. Brain 127, 791–800. https://doi.org/10.1093/BRAIN/AWH088.

Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Müller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatr. https://doi.org/10.1038/mp.2013.78.

Dickerson, B.C., Fenstermacher, E., Salat, D.H., Wolk, D.A., Maguire, R.P., Desikan, R., Pacheco, J., Quinn, B.T., Kouwe, A. Van der, Greve, D.N., Blacker, D., Albert, M.S., Killiany, R.J., Fischl, B., 2008. Detection of cortical thickness correlates of cognitive performance: reliability across MRI scan sessions, scanners, and field strengths. Neuroimage 39, 10. https://doi.org/10.1016/J.NEUROIMAGE.2007.08.042.

Dora, L., Agrawal, S., Panda, R., Abraham, A., 2017. State-of-the-Art methods for brain tissue segmentation: a review. IEEE Rev. Biomed. Eng. https://doi.org/10.1109/RBME.2017.2715350.

Holm-bonferroni method. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (Eds.), 2013. Encyclopedia of Systems Biology. Springer New York, New York, NY, p. 902. https://doi.org/10.1007/978-1-4419-9863-7_100629.

Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A., Gorgolewski, K.J., 2017. MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. PLoS One 12, e0184661. https://doi.org/10.1371/journal.pone.0184661.

Fischl, B., 2012. FreeSurfer. Neuroimage. https://doi.org/10.1016/j.neuroimage.2012.01.021.

Franke, K., Gaser, C., 2019. Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained? Front. Neurol. 789. https://doi.org/10.3389/FNEUR.2019.00789, 0.

Gibb, R., Kovalchuk, A., 2018. Brain Development. Neurobiol. Brain Behav. Dev. 3–27. https://doi.org/10.1016/B978-0-12-804036-2.00001-7.

Gilmore, A.D., Buser, N.J., Hanson, J.L., 2021. Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. Brain Informatics 2021 (81 8), 1–15. https://doi.org/10.1186/S40708-021-00128-2.

Haynes, L., Ip, A., Cho, I.Y.K., Dimond, D., Rohr, C.S., Bagshawe, M., Dewey, D., Lebel, C., Bray, S., 2020. Grey and white matter volumes in early childhood: a comparison of voxel-based morphometry pipelines. Dev. Cogn. Neurosci. 46, 100875 https://doi.org/10.1016/J.DCN.2020.100875.

Hibar, D.P., Westlye, L.T., Doan, N.T., Jahanshad, N., Cheung, J.W., Ching, C.R.K., Versace, A., Bilderbeck, A.C., Uhlmann, A., Mwangi, B., Krämer, B., Overs, B., Hartberg, C.B., Abe, C., Dima, D., Grotegerd, D., Sprooten, E., Ben, E., Jimenez, E., Howells, F.M., Delvecchio, G., Temmingh, H., Starke, J., Almeida, J.R.C., Goikolea, J.M., Houenou, J., Beard, L.M., Rauer, L., Abramovic, L., Bonnin, M., Ponteduro, M.F., Keil, M., Rive, M.M., Yao, N., Yalin, N., Najt, P., Rosa, P.G., Redlich, R., Trost, S., Hagenaars, S., Fears, S.C., Alonso-Lana, S., Van Erp, T.G.M., Nickson, T., Chaim-Avancini, T.M., Meier, T.B., Elvsashagen, T., Haukvik, U.K., Lee, W.H., Schene, A.H., Lloyd, A.J., Young, A.H., Nugent, A., Dale, A.M., Pfennig, A., McIntosh, A.M., Lafer, B., Baune, B.T., Ekman, C.J., Zarate, C.A., Bearden, C.E., Henry, C., Simhandl, C., McDonald, C., Bourne, C., Stein, D.J., Wolf, D.H., Cannon, D.M., Glahn, D.C., Veltman, D.J., Pomarol-Clotet, E., Vieta, E., Canales-Rodriguez, E.J., Nery, F.G., Duran, F.L.S., Busatto, G.F., Roberts, G., Pearlson, G.D., Goodwin, G.M., Kugel, H., Whalley, H.C., Ruhe, H.G., Soares, J.C., Fullerton, J.M., Rybakowski, J.K., Savitz, J., Chaim, K.T., Fatjó, Vilas, M., Soeiro-De-Souza, M.G., Boks, M.P., Zanetti, M.V., Otaduy, M.C.G., Schaufelberger, M.S., Alda, M., Ingvar, M., Phillips, M.L., Kempton, M.J., Bauer, M., Landén, M., Lawrence, N.S., Van Haren, N.E.M., Horn, N.R., Freimer, N.B., Gruber, O., Schofield, P.R., Mitchell, P.B., Kahn, R.S., Lenroot, R., Machado-Vieira, R., Ophoff, R.A., Sarró, S., Frangou, S., Satterthwaite, T.D., Hajek, T., Dannlowski, U., Malt, U.F., Arolt, V., Gattaz, W.F., Drevets, W.C., Caseras, X., Agartz, I., Thompson, P.M., Andreassen, O.A., 2018. Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Mol. Psychiatr. 23, 932. https://doi.org/10.1038/MP.2017.73.

Hutton, C., Vita, E. De, Ashburner, J., Deichmann, R., Turner, R., 2008. Voxel-based cortical thickness measurements in MRI. Neuroimage 40, 1701. https://doi.org/10.1016/J.NEUROIMAGE.2008.01.027.

Ibarretxe-Bilbao, N., Junque, C., Segura, B., Baggio, H.C., Marti, M.J., Valldeoriola, F., Bargallo, N., Tolosa, E., 2012. Progression of cortical thinning in early Parkinson's disease. Mov. Disord. 27, 1746–1753. https://doi.org/10.1002/MDS.25240.

Jubault, T., Gagnon, J.F., Karama, S., Ptito, A., Lafontaine, A.L., Evans, A.C., Monchi, O., 2011. Patterns of cortical thickness and surface area in early Parkinson's disease. Neuroimage 55, 462–467. https://doi.org/10.1016/J.NEUROIMAGE.2010.12.043.

Khundrakpam, B.S., Lewis, J.D., Kostopoulos, P., Carbonell, F., Evans, A.C., 2017. Cortical thickness abnormalities in autism spectrum disorders through late childhood, adolescence, and adulthood: a large-scale MRI study. Cerebr. Cortex 27, 1721–1731. https://doi.org/10.1093/CERCOR/BHX038.

Kl, N., Rp, W., Pm, T., P, S., D, R., T, D., M, G., Aw, T., Rm, B., 2007. Relationships between IQ and regional cortical gray matter thickness in healthy adults. Cerebr. Cortex 17, 2163–2171. https://doi.org/10.1093/CERCOR/BHL125.

LaMontagne, P., Benzinger, T., Morris, J., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., Raichle, M., Cruchaga, C., Marcus, D., 2019. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. medRxiv 2019. https://doi.org/10.1101/2019.12.13.19014902, 12.13.19014902.

Li, Q., Zhao, Y., Chen, Z., Long, J., Dai, J., Huang, X., Lui, S., Radua, J., Vieta, E., Kemp, G.J., Sweeney, J.A., Li, F., Gong, Q., 2019. 2019. Meta-analysis of cortical thickness abnormalities in medication-free patients with major depressive disorder. Neuropsychopharmacology 454 45, 703–712. https://doi.org/10.1038/s41386-019-0563-9.

Madan, C.R., 2018. Age differences in head motion and estimates of cortical morphology. PeerJ 2018, e5176. https://doi.org/10.7717/PEERJ.5176/SUPP-1.

Masouleh, S.K., Eickhoff, S.B., Zeighami, Y., Lewis, L.B., Dahnke, R., Gaser, C., Chouinard-Decorte, F., Lepage, C., Scholtens, L.H., Hoffstaedter, F., Glahn, D.C., Blangero, J., Evans, A.C., Genon, S., Valk, S.L., 2020. Influence of processing pipeline on cortical thickness measurement. Cerebr. Cortex 30, 5014–5027. https://doi.org/10.1093/CERCOR/BHAA097.

Ossenkoppele, R., Smith, R., Ohlsson, T., Strandberg, O., Mattsson, N., Insel, P.S., Palmqvist, S., Hansson, O., 2019. Associations between tau, Aβ, and cortical thickness with cognition in Alzheimer disease. Neurology 92, e601–e612. https://doi.org/10.1212/WNL.0000000000006875.

Pardoe, H.R., Kucharsky Hiess, R., Kuzniecky, R., 2016. Motion and morphometry in clinical and nonclinical populations. Neuroimage 135, 177–185. https://doi.org/10.1016/J.NEUROIMAGE.2016.05.005.

Reuter, M., Tisdall, M.D., Qureshi, A., Buckner, R.L., Kouwe, A.J.W. van der, Fischl, B., 2015. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. Neuroimage 107, 107. https://doi.org/10.1016/J.NEUROIMAGE.2014.12.006.

Seiger, R., Ganger, S., Kranz, G.S., Hahn, A., Lanzenberger, R., 2018. Cortical thickness estimations of FreeSurfer and the CAT12 toolbox in patients with Alzheimer's disease and healthy controls. J. Neuroimaging 28, 515–523. https://doi.org/10.1111/jon.12521.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428. https://doi.org/10.1037//0033-2909.86.2.420.

Sowell, E.R., Thompson, P.M., Leonard, C.M., Welcome, S.E., Kan, E., Toga, A.W., 2004. Longitudinal mapping of cortical thickness and brain growth in normal children. J. Neurosci. 24, 8223. https://doi.org/10.1523/JNEUROSCI.1798-04.2004.

Van Haren, N.E.M., Schnack, H.G., Cahn, W., Van Den Heuvel, M.P., Lepage, C., Collins, L., Evans, A.C., Hulshoff Pol, H.E., Kahn, R.S., 2011. Changes in cortical thickness during the course of illness in schizophrenia. Arch. Gen. Psychiatr. 68, 871–880. https://doi.org/10.1001/ARCHGENPSYCHIATRY.2011.88.

Van Rooij, D., Anagnostou, E., Arango, C., Auzias, G., Behrmann, M., Busatto, G.F., Calderoni, S., Daly, E., Deruelle, C., Di Martino, A., Dinstein, I., Duran, F.L.S., Durston, S., Ecker, C., Fair, D., Fedor, J., Fitzgerald, J., Freitag, C.M., Gallagher, L., Gori, I., Haar, S., Hoekstra, L., Jahanshad, N., Jalbrzikowski, M., Janssen, J., Lerch, J., Luna, B., Martinho, M.M., McGrath, J., Muratori, F., Murphy, C.M., Murphy, D.G.M., O'Hearn, K., Oranje, B., Parellada, M., Retico, A., Rosa, P., Rubia, K., Shook, D., Taylor, M., Thompson, P.M., Tosetti, M., Wallace, G.L., Zhou, F., Buitelaar, J.K., 2018. Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the ENIGMA ASD working group. Am. J. Psychiatr. 175, 359. https://doi.org/10.1176/APPI.AJP.2017.17010100.

Velázquez, J., Mateos, J., Pasaye, E.H., Barrios, F.A., Marquez-Flores, J.A., 2021. Cortical thickness estimation: a comparison of FreeSurfer and three voxel-based methods in a test-retest analysis and a clinical application. Brain Topogr. 34, 430–441. https://doi.org/10.1007/S10548-021-00852-2.

Vovk, U., Pernuš, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans. Med. Imag. https://doi.org/10.1109/TMI.2006.891486.

Walhovd, K.B., Fjell, A.M., Giedd, J., Dale, A.M., Brown, T.T., 2016. Through thick and thin: a need to reconcile contradictory results on trajectories in human cortical development. Cerebr. Cortex 27, 1–10. https://doi.org/10.1093/CERCOR/BHV301.

Wallace, G.L., Dankner, N., Kenworthy, L., Giedd, J.N., Martin, A., 2010. Age-related temporal and parietal cortical thinning in autism spectrum disorders. Brain 133, 3745. https://doi.org/10.1093/BRAIN/AWQ279.

Wilke, M., Holland, S.K., Altaye, M., Gaser, C., 2008. Template-O-Matic: a toolbox for creating customized pediatric templates. Neuroimage 41, 903–913. https://doi.org/10.1016/j.neuroimage.2008.02.056.

Yan, J., Cui, Y., Li, Q., Tian, L., Liu, B., Jiang, T., Zhang, D., Yan, H., 2019. Cortical thinning and flattening in schizophrenia and their unaffected parents. Neuropsychiatric Dis. Treat. 15, 935. https://doi.org/10.2147/NDT.S195134.

Yousaf, T., Dervenoulas, G., Politis, M., 2018. Advances in MRI methodology. Int. Rev. Neurobiol. 141, 31–76. https://doi.org/10.1016/BS.IRN.2018.08.008.