Physica Medica 64 (2019) 261–272

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Technical note

Evaluation of the intra- and inter-method agreement of brain MRI segmentation software packages: A comparison between SPM12 and FreeSurfer v6.0

L. Palumbo[a,*], P. Bosco[a], M.E. Fantacci[b], E. Ferrari[a,c], P. Oliva[d], G. Spera[a], A. Retico[a]

[a] National Institute for Nuclear Physics (INFN), Pisa Division, Pisa, Italy
[b] University of Pisa, Physics Department, Pisa, Italy
[c] Scuola Normale Superiore, Pisa, Italy
[d] University of Sassari and INFN Cagliari Division, Italy

ARTICLE INFO

Keywords:
Brain MRI
Segmentation
Repeatability
Reproducibility

ABSTRACT

Purpose: The lack of inter-method agreement can produce inconsistent results in neuroimaging studies. We evaluated the intra-method repeatability and the inter-method reproducibility of two widely-used automatic segmentation methods for brain MRI: the FreeSurfer (FS) and the Statistical Parametric Mapping (SPM) software packages.
Methods: We segmented the gray matter (GM), the white matter (WM) and subcortical structures in test-retest MRI data of healthy volunteers from Kirby-21 and OASIS datasets. We used Pearson's correlation (r), Bland-Altman plot and Dice index to study intra-method repeatability and inter-method reproducibility. In order to test whether different processing methods affect the results of a neuroimaging-based group study, we carried out a statistical comparison between male and female volume measures.
Results: A high correlation was found between test-retest volume measures for both SPM (r in the 0.98–0.99 range) and FS (r in the 0.95–0.99 range). A non-null bias between test-retest FS volumes was detected for GM and WM in the OASIS dataset. The inter-method reproducibility analysis measured volume correlation values in the 0.72–0.98 range and the overlap between the segmented structures assessed by the Dice index was in the 0.76–0.83 range. SPM systematically provided significantly greater GM volumes and lower WM and subcortical volumes with respect to FS. In the male vs. female brain volume comparisons, inconsistencies arose for the OASIS dataset, where the gender-related differences appear subtler with respect to the Kirby dataset.
Conclusions: The inter-method reproducibility should be evaluated before interpreting the results of neuroimaging studies.

1. Introduction

In the last 20 years, magnetic resonance imaging (MRI) has allowed to non-invasively study brain structure at high resolutions. In neuroscience research, segmenting a brain MRI into different structures is a widely used pre-processing step [1,2]. Even though the manual segmentation is considered the gold standard [3], it is operator-dependent, laborious, and time-consuming. Automated segmentation tools are used to segment the brain regions in a reasonable amount of time and thus they are essential for investigating large datasets. These automatic methods have been used in a large number of research studies on psychiatric and neurological disorders. However, the reliability of their measurements is a matter of debate [4], in terms of both reproducibility and repeatability. The reproducibility is the measure of agreement between the results obtained with the two methods on the same scan of a subject. The repeatability is the degree of agreement between brain volumes obtained with the same method on an identical subject in two subsequent acquisitions (test-retest analysis).

Several studies [4–7] focused on the test-retest repeatability assessment of automatic segmentation methods. For example, Chard et al. [5] investigated the repeatability of gray matter (GM) and white matter (WM) volumes obtained with the Statistical Parametric Mapping (SPM) software package, showing that SPM99 was reliable to segment GM and WM, with Pearson's correlations of 0.908 and 0.895, respectively. In addition, they showed that the inhomogeneity correction improves the repeatability. Several studies reported high inter-method

[*] Corresponding author at: National Institute for Nuclear Physics (INFN), Pisa Division, Largo Bruno Pontecorvo 3,56127, Pisa, Italy. Tel.: 0502214879.
E-mail address: letizia.palumbo@pi.infn.it (L. Palumbo).

https://doi.org/10.1016/j.ejmp.2019.07.016
Received 19 March 2019; Received in revised form 12 June 2019; Accepted 19 July 2019
Available online 05 August 2019

reproducibility between automatic segmentation methods, but some of them highlight the presence of method-specific biases [6–12]. In fact, processing methods can systematically either underestimate or overestimate the volumes of different brain regions with respect to a reference volume, defined for instance by manual segmentation. Relevant differences between SPM and FreeSurfer (FS) were shown by Katuwal et al. [8] in the estimated GM and WM volumes. In that case, the volumes obtained with SPM8 were closer to manual segmentations than those obtained with FS v5.3; in particular, SPM8 overestimated GM and WM volumes with respect to FS v5.3, by 40% and 26%, respectively. In the study by Wenger et al. [9], FS v5.3 was shown to overestimate the hippocampal volume with respect to manual segmentation. However, even though Kazemi et al. [10] showed the superior accuracy of SPM8 in segmenting brain tissues with respect to other automated methods, they also pointed out that it is not suitable for the segmentation of subcortical structures. Perlaki et al. [13] showed that FS v4.5/v5.3 was less accurate in the segmentation of putamen with respect to FSL-FIRST. The discrepancies in volume estimates between different segmentation algorithms could reside in the different image processing algorithms implemented (e.g. the use of different templates and image registration methods) [8].

The reliability of automatic segmentation methods and the comparison between automated segmentation software packages has been evaluated in various studies [14–16]. A summary of structural MRI studies on the repeatability of automatic segmentation software is reported in Table 1, together with the investigated brain regions and used methods.

In this work we provide a direct and appropriate comparison between the latest versions of the software packages of two popular segmentation methods, Statistical Parametric Mapping (v12) [17,18] and FreeSurfer (v6.0) [19,20], using two publicly available datasets, namely Kirby [21] and OASIS [22].

The aim of this work is to analyze the intra-method repeatability, the inter-method reproducibility and the possible systematic bias in the volume estimates generated by the two automatic segmentation software, focusing on six regions of interest (ROIs): two global measures (GM and WM) and four subcortical structures (hippocampus, putamen, caudate and brainstem). The reliability of each method and the quantification of any discrepancy are relevant for the correct interpretation of the results of longitudinal studies and for quantitative considerations in *meta*-analyses.

In order to test how the use of different segmentation algorithms could affect the results of volume group comparisons, we analyzed the differences in brain volume measures between male and female subgroups in both OASIS and Kirby data samples, using SPM12 and FS v6.0.

2. Materials and methods

2.1. Data samples

We examined two publicly available data samples: the Kirby-21 (Kirby) dataset [21,23], and the OASIS dataset [22,24].

The Kirby dataset consists of 3D T1-weighted images of 21 healthy volunteers (11 males and 10 females; age: 32 ± 9 years) that were acquired using a 3 T MRI scanner. The acquisition protocol included whole-brain high-resolution anatomical 3D images (MPRAGE sequence, TE/TR 6.7/3.1 ms, 1 × 1 × 1.2 mm voxel size, flip angle 8°). For repeatability studies, MRI images were acquired twice: after the first session, each subject left the scan room for a short break and then he/she was repositioned in the scanner for an identical session with the same acquisition parameters [21].

The OASIS dataset consists of 3D T1-weighted images of 20 healthy volunteers (8 males and 12 females; age: 23.4 ± 3.9 years), acquired using a 1.5 T MRI scanner. The acquisition protocol included a whole-brain high-resolution anatomical scan (MPRAGE sequence, TE/TR 4.0/9.7 ms, 1x1x1 mm voxel size, flip angle 10°) [16,25]. The 20 subjects were scanned twice with a time delay in the range of 1–89 days (mean delay of about 21 days) to enable repeatability studies [22,25].

An example of anatomical scans of two datasets is shown in Fig. 1.

The test-retest analysis allows estimating a tool's reliability in measuring a given quantity in the successive repetition of the measurement under the same conditions. MRI data may be affected by confounding factors that may vary between scans, like the field of view, or patient positioning. In addition, there are effects such as hydration levels that cause possible day-to-day variations in the brain structures [4,26].

2.2. Segmentation and volume measures

We estimated the volumes of six different brain tissues: GM, WM and four subcortical structures (hippocampus, putamen, caudate and brainstem), using two processing methods: FS v6.0 and SPM12. Both FS and SPM are a set of tools and algorithms to extract measures from neuroimaging data for the study of the human brain in healthy and pathological conditions.

FreeSurfer is used as a pre-processing workflow for structural MRI data (recon-all analysis pipeline) [19], which performs all cortical reconstruction through 31 processing steps. The FS pipeline is shown in Fig. 2(a).

To carry out brain tissue segmentation, FS takes advantage of a lot of information, e.g. image intensities, global position within the brain and position relative to neighboring brain structures. Then, it uses

Table 1
Structural MRI studies on the repeatability assessment of automatic segmentation methods.

Reference	Brain regions	Methods
Chard et al. [5]	Gray Matter, White Matter	SPM99
Jovicich et al. [15]	Hippocampus, Thalamus, Caudate, Putamen, Pallidum, Amygdala, Lateral Ventricles, Inferior Lateral Ventricles, Intracranial	FS 4.0
Katuwal et al. [8]	Gray Matter, White Matter, Cerebrospinal fluid, Total intracranial volume	SPM8, FSL 5.0.4, FS 5.3
Kazemi et al. [10]	Gray Matter, White Matter, Cerebrospinal Fluid	SPM8, FSL 4.1, Brainsuite
Maclaren et al. [4]	Hippocampus, Lateral ventricles, Amygdala, Putamen, Pallidum, Caudate, Thalamus, Cerebral White Matter	FS 5.1
Morey et al. [28]	Amygdala, Brain Stem, Hippocampus, Lateral Ventricles, Nucleus Accumbens, Caudate, Putamen, Pallidum, Thalamus	FSL-FIRST 1.2, FS 4.5
Ochs et al. [7]	Total intracranial volume, Whole brain parenchyma, Cortical gray matter, Lateral ventricle, Inferior lateral ventricle, 3rd ventricle, 4th ventricle, Total cerebrospinal fluid, Caudate, Putamen, Pallidum, Thalamus, Amygdala, Hippocampus, Ventral diencephalon, Cerebellar white matter, Cerebellar gray matter, Cerebellum, Brain Stem	FS 5.3, NeuroQuant. 1.4
Perlaki et al. [13]	Caudate, Putamen	FSL 5.0.7, FS 4.5, FS 5.3
Tae et al. [14]	Hippocampus	FS 3.0.4, IBASPM (SPM2)
Wenger et al. [9]	Hippocampus	FS 5.3

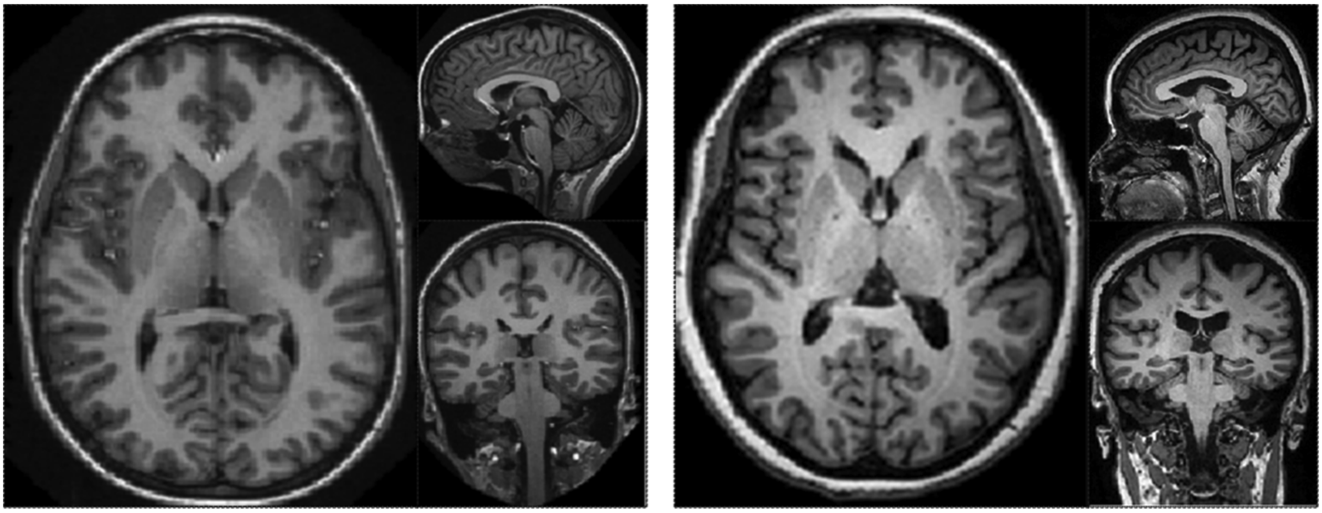

Fig. 1. Original axial, sagittal, coronal view of the 3D brain MRI data of two datasets OASIS (left) and Kirby-21 (right).

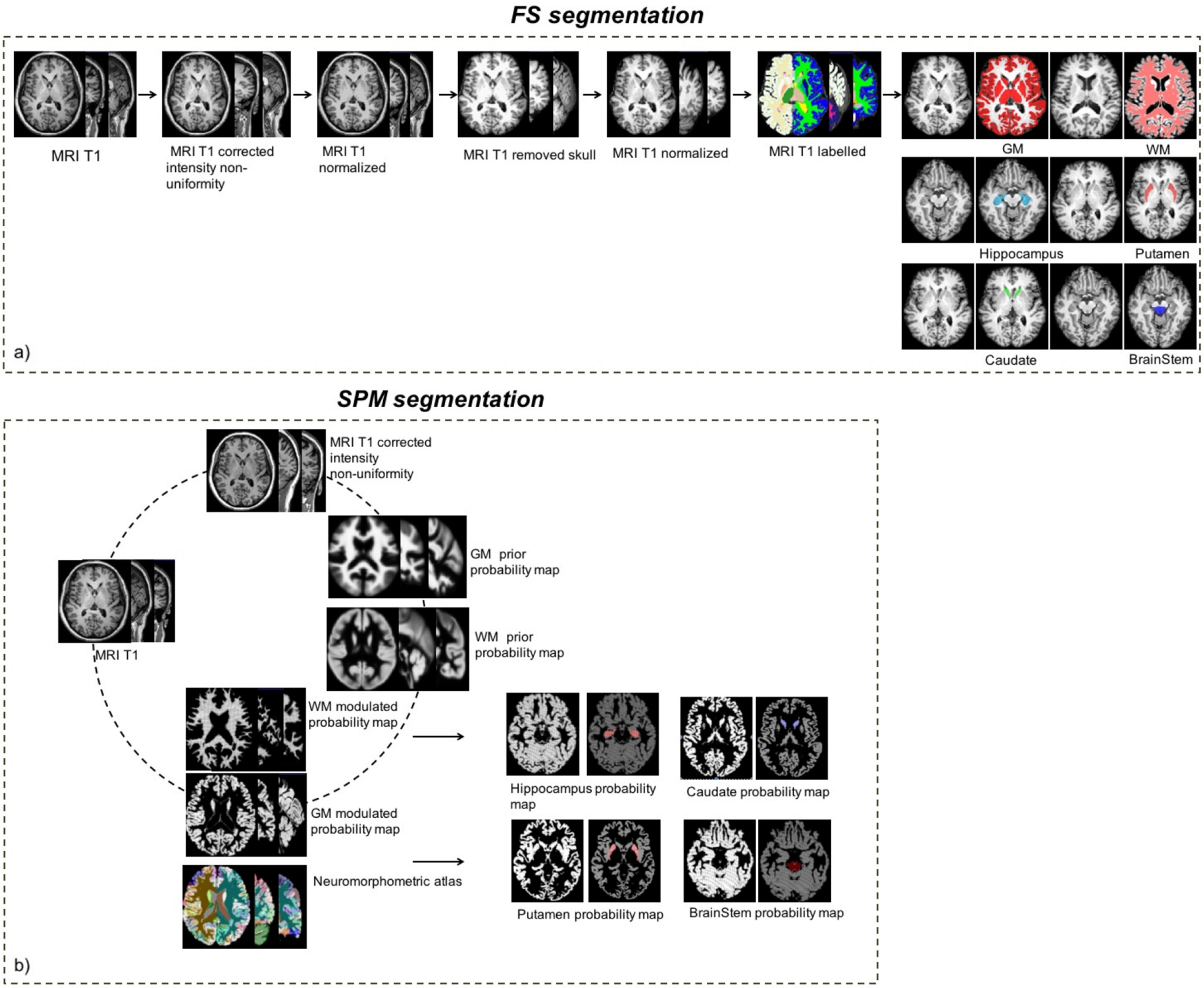

Fig. 2. (a) FreeSurfer analysis pipeline showing the main steps of interest from motion correction to automated subcortical segmentation; (b) SPM preprocessing pipeline, including also the use of the Neuromorphometric atlas.

probabilistic atlas in which coordinates have anatomical meaning and the Markov random field (MRF) model is used to find local spatial relationships between labeled structures. FS implements a model based on a mixture of a small number of Gaussians for each structure for each point in the space and maximum posterior estimate of the model parameters to assign one of the 37 ROI labels to each voxel [27]. FreeSurfer allows for manual editing of segmentation results; however, it is an extremely time-consuming procedure and McCarthy et al. [28] did not find significant differences between segmented volumes with and without manual corrections. The volumes of the ROIs used in this study have been extracted from the aseg.stats FS output [19].

SPM allows the segmentation of brain structural data in GM, WM and cerebrospinal fluid (CSF), and implements a variety of processing algorithms and statistical functions to make voxel-wise group analyses. The SPM analysis pipeline is shown in Fig. 2(b). Before segmenting the brain structures it needs to align the image on the anterior commissure to avoid possible segmentation algorithm failure. The *spm_prepoc_run* script allows the segmentation into different tissue classes, using a modified Gaussian Mixture Model. It includes three steps: 1) non-uniformity intensity correction; 2) registration to the tissue probability maps representing the prior probability of different tissue classes; 3) posterior probability computation using Bayes model to combine the prior probability with the tissue type probabilities derived from voxel intensities. It is a circular process that includes classification, bias correction and registration treads [29]. To mitigate the effect of partial volume it is appropriate to set the number of Gaussians to 2 for each tissue since a voxel could contain a signal from a number of different tissues [18]. To preserve the total amount of GM and WM the modulation operation is applied, in order to correct for regional enlargement/shrinkage of the volumes during spatial normalization, the warped images are multiplied, voxel-by-voxel, with the relative volumes of tissue (i.e. the Jacobian determinants of the deformations). To segment the subcortical structures for this study (hippocampus, putamen, caudate and brainstem) an extension of SPM has been considered, implementing the atlas Neuromorphometric labels [17], which can be used to mask GM and WM probability maps. Brain tissues volumes were obtained by using the *spm_get_volumes* script [8].

We extracted the brain tissue volumes (GM, WM) and subcortical structures volumes (hippocampus, putamen, caudate, brainstem) obtained with SPM and FS for all the subjects of both data samples, and we used these values to carry out the repeatability and reproducibility analyses.

2.3. Repeatability, reproducibility and bias measures

Firstly, we quantified the intra-method repeatability of each method in estimating brain structures volume through a test-retest analysis. Secondly, we evaluated the inter-method reproducibility of estimated volumes comparing the two processing methods. These analyses were conducted in parallel for the two available data samples.

To evaluate the intra-method repeatability, we first computed the Pearson's correlation between the volumes obtained for each brain structure of each subject in the test and retest scans; then, to quantify the agreement we implemented the Bland-Altman (BA) plot representation [30], reporting the percentage differences between test and retest measures for each subject. In this plot, for each pair the percentage difference of the two measurements $\frac{V_{scan} - V_{rescan}}{\left(\frac{V_{scan} + V_{rescan}}{2}\right)}*100$ is reported as a function of the average measured volume $\frac{V_{scan} + V_{rescan}}{2}$, together with the mean (d) of this average percentage difference and the limits of 95% confidence interval (C.I.) agreement.

We checked the normality of the distribution of the differences with the Shapiro-Wilk test. We used Bland-Altman plots to detect possible systematic biases between test and retest scans. The test-retest measured volumes can be considered equal if the percentage difference is null within the limits of agreements.

To evaluate the inter-method reproducibility, we used the brain structures volumes obtained in the segmentation of the first set of scans, for each dataset separately. We computed Pearson's correlation between the volume measures obtained with the two different software. This analysis was performed for both datasets. In addition, to estimate possible systematic differences between the segmented volumes with the two different methods, we implemented the BA plot representation.

2.4. A practical example of between-group comparison

To investigate whether the use of the two different preprocessing pipelines (SPM12 and FS v6.0) has a direct impact on the results of a neuroscience study, we compared the brain volume measures obtained with each software for the male and female subsamples, in order to reveal gender-related volume differences [11,16,31–33].

For each method, we tested male-female differences in all brain volumes on both datasets with *t*-test and calculated the effect sizes, Cohen's d [34] as:

$$d = \frac{M_{male} - M_{female}}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{(n_{male} - 1)\sigma^2_{male} + (n_{female} - 1)\sigma^2_{female}}{(n_{male} + n_{female} - 2)}}$$

Possible variations in the shape of the segmented regions have been quantified in terms of the Dice (D) similarity index, which is an overlap measure between two binary images, defined as:

$$D = \frac{2|A \cap B|}{|A| + |B|}$$

and ranging from 0 to 1. A Dice index equal to 0 means that there is no overlap, whereas D equals to 1 means that the overlap is perfect [35]. To compare SPM and FS segmented volumes, the SPM tissues probability maps were binarized using a 0.5 threshold. We visually verified that brain masks derived using such thresholds were more accurately defined than by using other values [36].

By contrast, FS segmented ROIs are already provided as binary masks. The spatial comparison has been carried out in the native space of the images. Thus, it was necessary to transform back to the native space the segmented ROI maps for both methods. For FS we used the library function '***mri_label2vol***', which converts a label mask into a mask in the native space [19]. For SPM we transformed back to the native space the brain masks using the '***spm_normalize_to_write***' and the ***'spm_realign'*** scripts [17].

3. Results

3.1. Intra-method repeatability

Pearson's correlations between the volumes obtained on test and retest MRI data are reported in Table 2 for both Kirby and OASIS data samples. SPM showed high Pearson's correlation values on both data samples (r in the 0.98–0.99 range for GM, MW and in the four considered subcortical regions). FS showed high values of r on the Kirby dataset (r in the 0.98–0.99 range for GM and MW, r in the 0.95–0.99 range for the four subcortical regions). We also reported in Table 2 the average percentage differences (d) between test-retest volumes, their standard deviation (s) over the sample and the limits of agreement corresponding to the 95% C.I. in the measure of d for both SPM and FS on the Kirby and OASIS data samples.

The Bland-Altman plots are shown in Figs. 3 and 4 for GM and WM. The BA plots for the subcortical volumes are provided in the Supplementary Material.

The mean percent differences for SPM are always consistent with zero, showing that the volume estimates are repeatable, whereas for FS

Table 2
Intra-method repeatability measurements for SPM and FS evaluated on the Kirby and OASIS data samples: Pearson's correlation and Bland-Altman plot parameters, e.g. mean (d) and standard deviation (s) of percent difference and limits of agreement corresponding to the 95% C.I. in volume difference.

Kirby dataset								
	SPM				FS			
Brain region	Pearson's correlation	d (%)	s (%)	Limits of agreement	Pearson's correlation	d (%)	s (%)	Limits of agreement
GM	0.99	0	1.1	−2.2 to 2.1	0.98	1.0	2.0	−2.9 to 5.0
WM	0.99	0	0.8	−1.6 to 1.5	0.98	−0.7	1.8	−4.2 to 2.8
Hippocampus	0.99	−0.2	0.7	−1.5 to 1.2	0.95	−1	3	−7 to 6
Putamen	0.99	−0.3	1.7	−3.6 to 3.0	0.99	−0.4	1.7	−3.9 to 3.0
Caudate	0.99	−0.1	0.8	−1.8 to 1.5	0.99	0	1.5	−2.8 to 2.9
Brainstem	0.99	−0.4	0.7	−1.7 to 1.0	0.99	0.2	1.5	−2.8 to 3.2
OASIS dataset								
	SPM				FS			
Brain region	Pearson's correlation	d (%)	s (%)	Limits of agreement	Pearson's correlation	d (%)	s (%)	Limits of agreement
GM	0.99	0.1	1.4	−2.7 to 2.9	0.99	4.6	1.3	2.1–7.1
WM	0.99	−0.5	1.3	−2.9 to 1.9	0.99	−2.8	1.2	−5.1 to −0.5
Hippocampus	0.98	−1.3	0.8	−2.8 to 0.2	0.96	−2.7	2.0	−6.6 to 1.2
Putamen	0.98	3.1	3.0	−2.7 to 8.9	0.98	1.0	2.6	−4.0 to 6.1
Caudate	0.99	0.7	1.3	−1.9 to 3.2	0.99	1.2	2.1	−2.9 to 5.2
Brainstem	0.99	−0.1	1.7	−3.5 to 3.3	0.99	−0.5	2.3	−5.0 to 3.9

this consideration is true in all comparisons except for the GM and WM volumes obtained on the OASIS database. For these quantities the test measures are systematically higher, d = (4.6 ± 1.3) % and lower d = (−2.8 ± 1.2) % than the retest measures, respectively.

3.2. Inter-method reproducibility

The inter-method reproducibility analysis has been conducted on the measures obtained from the first scan for each data sample. Bar

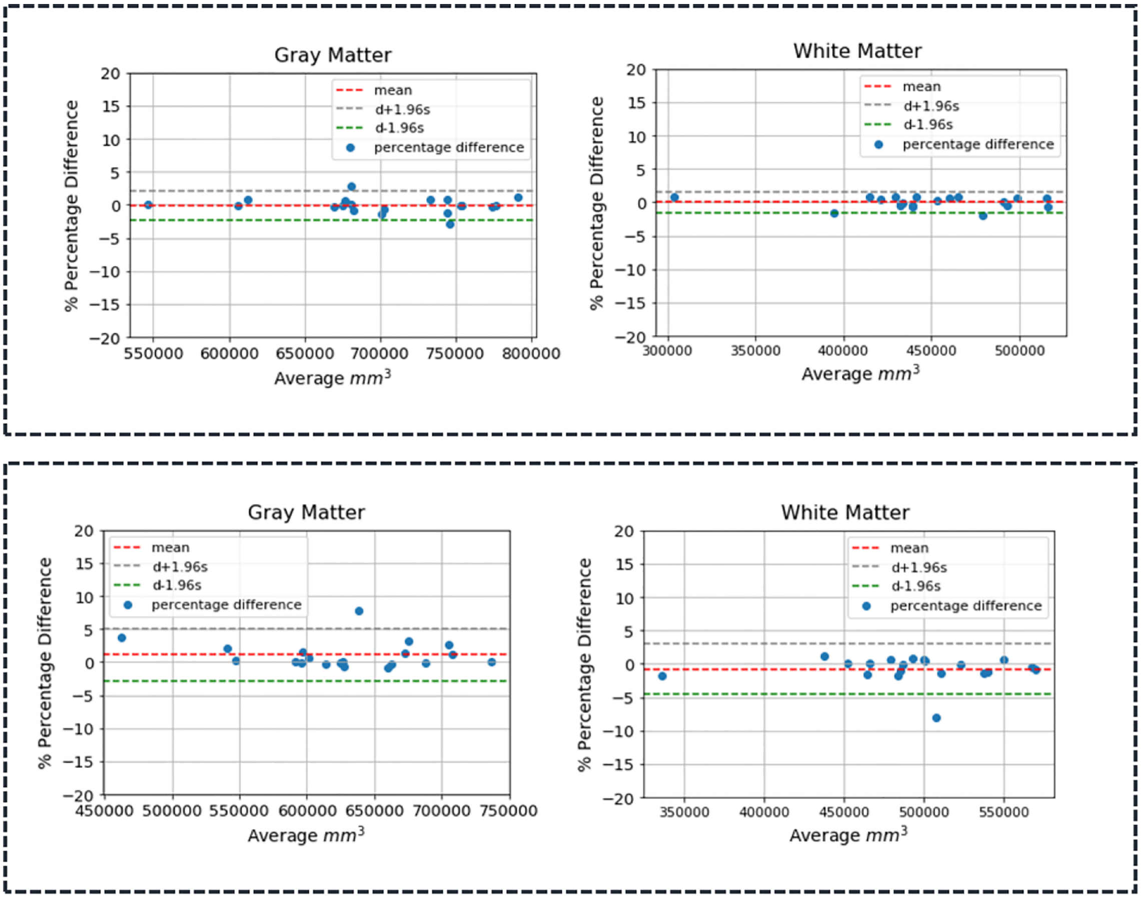

Fig. 3. Bland-Altman plots of segmented GM and WM by SPM (above) and by FS (below) in scan-rescan analysis on the Kirby-21 dataset. Bland-Altman plots of the four segmented subcortical structures are shown in the Supplementary Material.

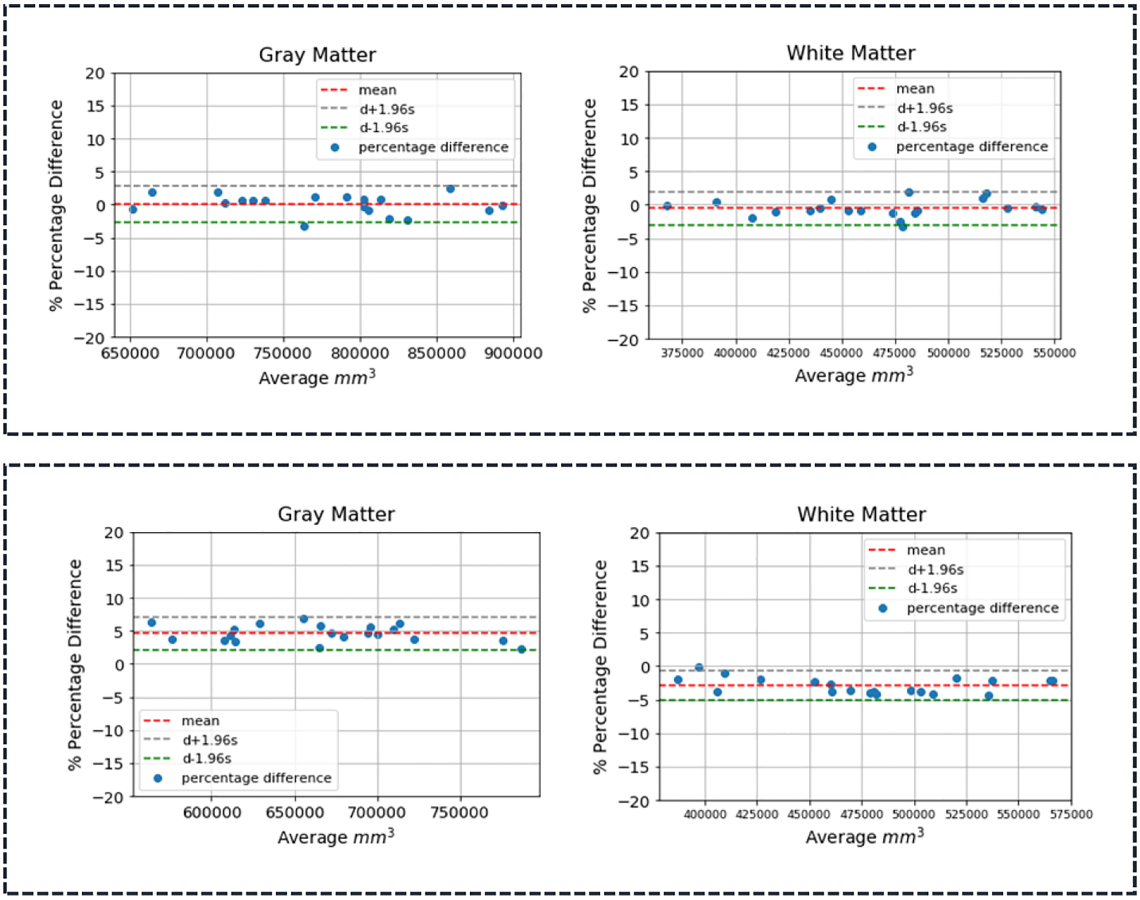

Fig. 4. Bland-Altman plots of segmented GM and WM by SPM (above) and by FS (below) in scan-rescan analysis on the OASIS dataset. Bland-Altman plots of the four segmented subcortical structures are shown in the Supplementary Material.

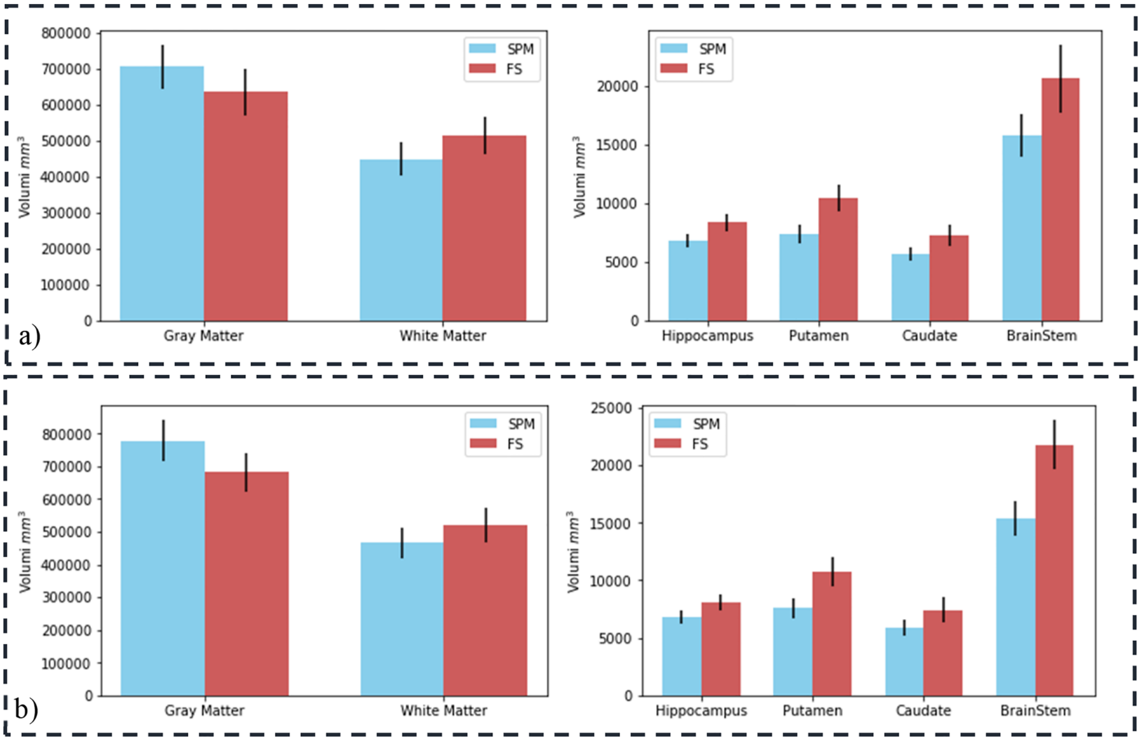

Fig. 5. Bar-plots of segmented brain tissues volumes and subcortical structures (a) on the Kirby-21 dataset, (b) on OASIS dataset.

Table 3
Inter-method reproducibility measurements between SPM and FS evaluated on the Kirby and OASIS data samples: Pearson's correlation; Bland-Altman parameters e.g. mean (d) and standard deviation (s) of percent difference and limits of agreement corresponding to the 95% C.I. in volume difference; Dice indices (D).

Kirby dataset					
Brain region	Pearson's correlation	d (%)	s (%)	Limits of agreement	D
GM	0.94	11	3	4–17	0.80 ± 0.03
WM	0.98	−13	2	−10 to −17	0.81 ± 0.03
Hippocampus	0.72	−21	6	−34 to −8	0.78 ± 0.03
Putamen	0.86	−34	5	−24 to −45	0.76 ± 0.04
Caudate	0.80	−23	8	−39 to −8	0.80 ± 0.02
Brainstem	0.95	−26	5	−36 to −17	0.84 ± 0.02

OASIS dataset					
Brain region	Pearson's correlation	d (%)	s (%)	Limits of agreement	D
GM	0.96	13	3	9–18	0.83 ± 0.01
WM	0.95	−11	3	−17 to − 5	0.83 ± 0.02
Hippocampus	0.78	−17	5	−28 to − 7	0.78 ± 0.02
Putamen	0.84	−34	6	−47 to − 22	0.77 ± 0.04
Caudate	0.93	−23	6	−35 to − 11	0.79 ± 0.02
Brainstem	0.95	−34	3	−40 to − 29	0.80 ± 0.02

plots of the volumes for the six ROIs are shown in Fig. 5(a) and (b) for Kirby and OASIS data samples, respectively. In Table 3 we report the Pearson's correlations between SPM and FS measures, as well as the percentage differences (d) averaged over the subjects, their standard deviations (s) and the 95% C.I. limits on d.

The Bland-Altman plots obtained with the measures of the Kirby and OASIS data samples are reported in Figs. 6 and 7 for GM and WM, respectively. The BA plots for the subcortical volumes are provided in the Supplementary Material. The Pearson's correlation coefficients are above 0.8 for all measures on both Kirby and OASIS data samples, except for the hippocampus (r = 0.72 on Kirby and r = 0.78 on OASIS). The Bland-Altman plots show systematic biases: SPM provides a significantly greater volume of GM and significantly lower volumes of WM and the subcortical ROIs with respect to FS.

These findings are consistently detected on both Kirby and OASIS data samples. Dice's similarity coefficients between SPM and FS segmentations for each brain structure are reported in Table 3. Dice values are in the 0.76–0.83 range. Consistently on both data samples, the smaller structures (hippocampus, putamen and caudate) show the worst overlap, generally below 0.8, whereas brain tissues (GM and WM) and the brainstem show Dice values generally above 0.8. The modest overlap of segmented structures ($D < 0.8$) occurs in particular in ROIs with poor contrast or difficulties in the boundary definition between GM and WM matter, such as those reported in Fig. 8 (putamen, caudate and hippocampus). Fig. 9 shows the overlay of the GM masks obtained with SPM and FS, where the difference in the definition of the brain regions belonging to the GM is clearly visible. For example, the thalamus is not included in the GM tissue identified by SPM. We reported the overlay of the ROI masks obtained by SPM and FS, for the worst cases of the Kirby and the OASIS data samples, respectively in Figs. 10 and 11.

3.3. Consistency test of male vs. female volume comparison

The two-sample *t*-test between volume measures of the male and female subgroups obtained with SPM12 and FS v6.0 consistently revealed significant gender-related differences on the Kirby data sample. In that case, the Cohen's d effect sizes [34] range from 0.76 (large) to 0.84 (huge) for both FS and SPM (see Table 4 and Fig. 12a). By contrast, gender-related differences appear subtler on the OASIS data sample, and no statistically significant differences are actually detected for the global WM and for the hippocampus, caudate and brainstem volume measures, consistently for SPM and FS (see Table 4 and Fig. 12b). Among the subcortical structures, only the putamen volume is found to be significantly larger in the male cohort and with very large Cohen's d values consistently for both segmentation methods. A relevant inconsistency between the two different segmentation methods was found for the GM volume. In fact, a significantly larger GM volume is found for the male cohort with a large Cohen's d according to FS measures, while the analysis of the SPM GM measure does not confirm this finding.

4. Discussion

We examined the intra- and inter-method agreement of two popular automatic brain MRI segmentation tools (SPM and FS), using the publicly available Kirby and OASIS datasets, containing high-resolution structural MRI images of healthy subjects. The choice of the methods and of the data samples makes this study reproducible and extendable by other researchers.

Although extremely high Pearson's correlation coefficients between test and retest measures have been found for both SPM and FS, the BA

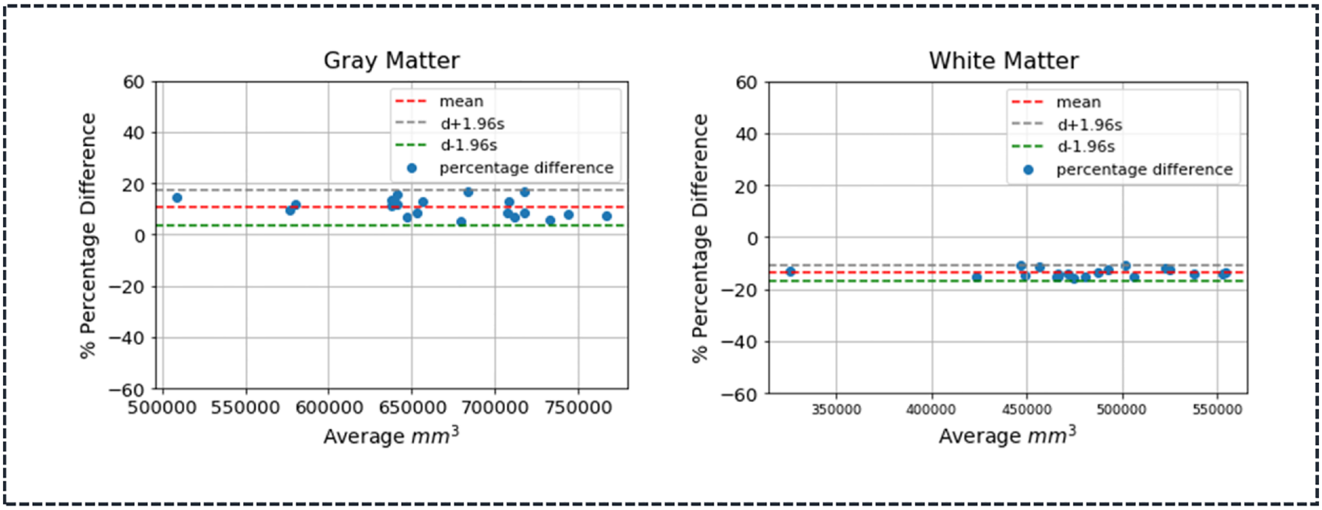

Fig. 6. Bland-Altman plots of segmented GM and WM by FS and SPM in the inter-method reproducibility analysis on the Kirby-21 dataset. Bland-Altman plots of the four segmented subcortical structures are shown in the Supplementary Material.

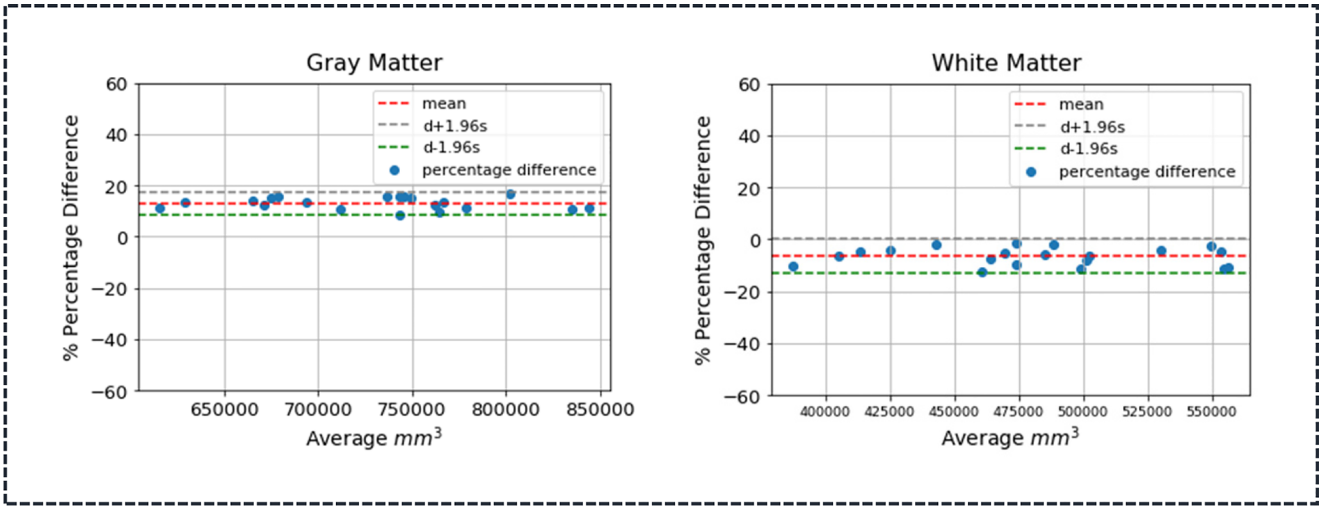

Fig. 7. Bland-Altman plots of segmented GM and WM by FS and SPM in the inter-method reproducibility analysis on the OASIS dataset. Bland-Altman plots of the four segmented subcortical structures are shown in the Supplementary Material.

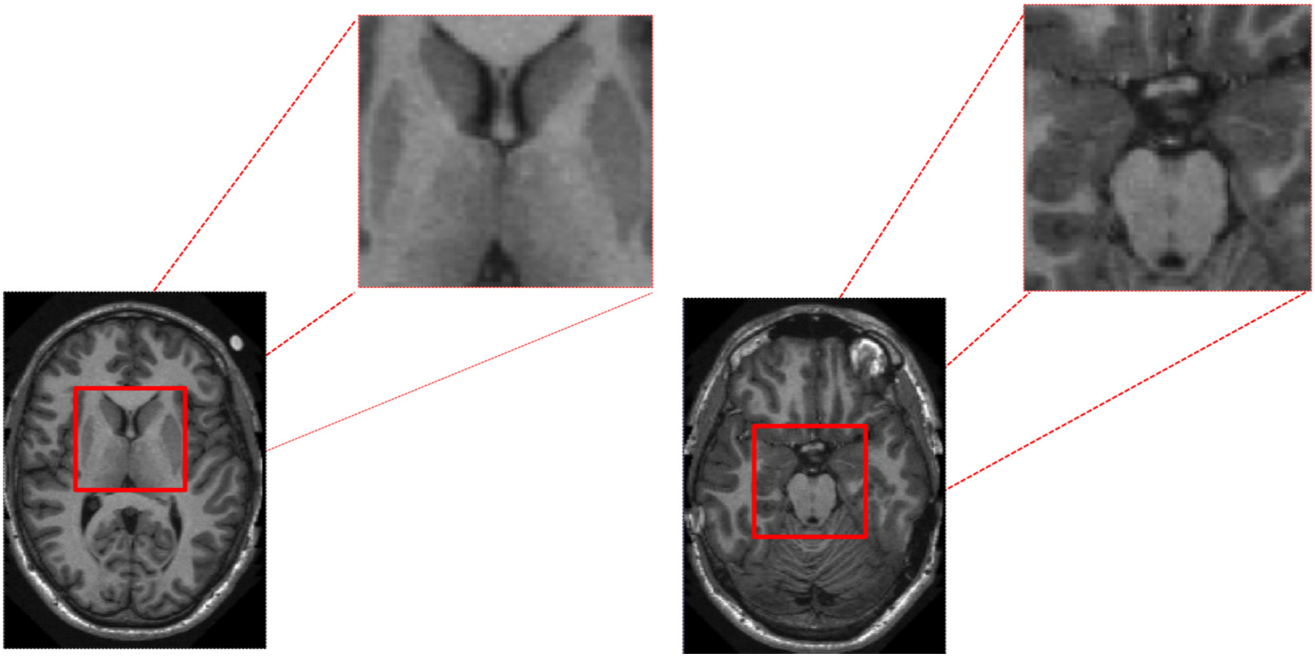

Fig. 8. Highlight of the poor contrast and difficult of boundary definition for some subcortical anatomical structures for one example subject: zoom on the putamen and caudate area (left); zoom on the hippocampus area (right).

plots revealed very good repeatability of both methods only on the Kirby dataset. By contrast, on the OASIS data sample, the BA analysis detected non-null biases for the FS measures: for example, for GM and WM d = 4.6% and d = −2.7% respectively. On the contrary, SPM showed no or smaller bias on them (d = 0.1% and d = −0.5% for GM and WM respectively).

The OASIS dataset presents several differences with respect to the Kirby dataset. First of all, there is a significant average time between test-retest is about three weeks (while Kirby retest acquisitions are made in the same day). However, a reduction of the GM close to 5% is not likely in this time-lapse, and a simultaneous increase of the WM close to 3% is definitely unrealistic. Another difference between the two datasets is the field strength (3 T for Kirby and 1.5 for OASIS). Regarding this point, literature reports consistency in volume estimation at 1.5 T and 3 T, for both FS and SPM [37]. Eventually, it is not possible to exclude the existence of differences in the populations underlying the two datasets (for example, the mean age in Kirby is 32 ± 9 and 23.4 ± 3.9 in OASIS). However, the subjects are healthy volunteers in both datasets and a significant difference in brain volumes is not expected.

These considerations lead us to conclude that SPM is more robust with respect to any variation that may have occurred between test and retest scans on that data sample.

It is worth noticing that on both data samples the standard deviations of the test-retest percent volume differences are generally higher when using FS, indicating that small differences in patient's orientations due to repositioning may have a greater effect on FS than on SPM measures.

The inter-method reproducibility analysis revealed discrepancies between the volumes calculated by SPM and FS. These discrepancies

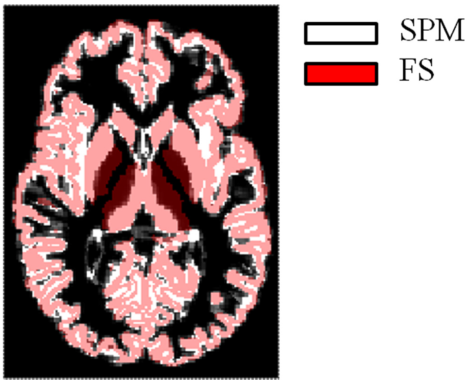

Fig. 9. GM ROI masks obtained with SPM (white) and FS (red). The difference in the definition of brain regions are clearly visible in the image (white: SPM; red: FS; pink: common regions). For example, the thalamus is not included in the GM tissue identified by SPM.

are visible both in the BA plots and in terms of the Dice indices. Larger values have been systematically found by FS for the WM and the four subcortical ROIs with respect to SPM. By contrast, the GM volume estimated by FS is lower with respect to the one estimated by SPM. Of note, one recently published study [38] found that SPM12 overestimates the segmented volumes, in line with our results. Indeed, their results suggest that FreeSurfer cortical thicknesses were lower compared to SPM12 values. These differences in the definition of the boundary between GM and WM can be due to the implementation of different segmentation algorithms in SPM and FS, including the adoption of different reference atlases (ICBM-452 [35] T1 brain atlas in SPM12 [17] and MNI 305 [39] in FS).

The Dice indices highlighted an overlap between the ROIs segmented with the two methods of about 80%, with lower values in case of subcortical ROIs. This result, which is consistently found on both data samples, can be understood in terms of the larger surface-to-volume ratio of smaller structures. In fact, smaller structures are more prone to segmentation discrepancies due to the different definitions of

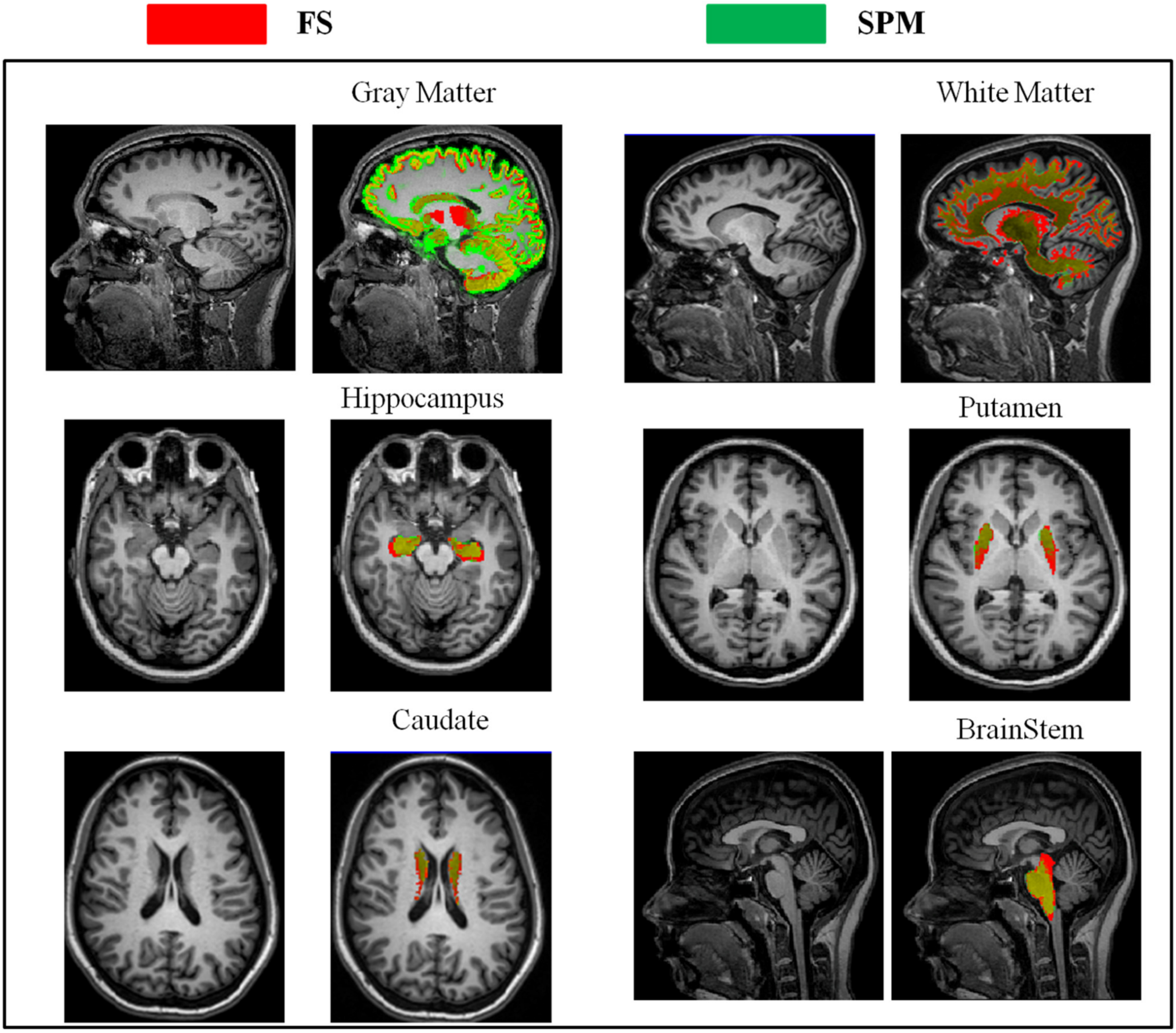

Fig. 10. Overlay of segmented ROIs by SPM and by FS onto a single subject anatomical image in the native space for the worst case of the Kirby-21 data sample. GM and WM are visible on the first line; hippocampus and caudate on the second and third lines, respectively.

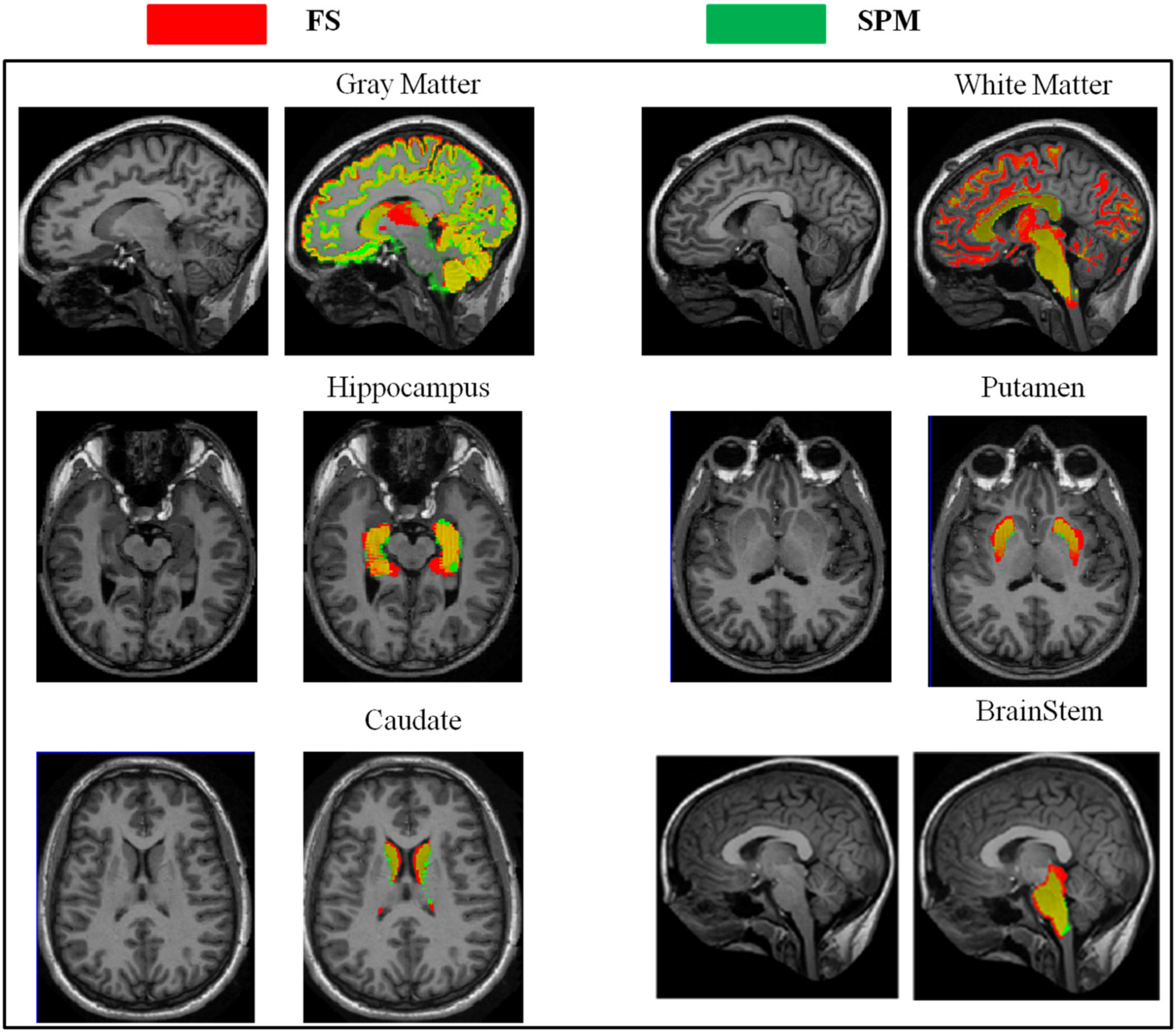

Fig. 11. Overlay of segmented ROIs by SPM and by FS onto a single subject anatomical image in the native space for the worst case of the OASIS data sample. GM and WM are visible on the first line; hippocampus and caudate on the second and third lines, respectively.

the boundary between GM and WM by the two algorithms. Additional sources of discrepancy in segmenting subcortical structures are provided by the arbitrariness in some cases in defining their boundaries with respect to surrounding structures, as the case of the hippocampus and the confining amygdala.

To test for possible inconsistencies in group comparison results due to the choice of different segmentation algorithms, we set up a two-group comparison between the only subgroups available in this study, i.e. the male and female cohorts. According to numerous studies [11,16,31–33], males are expected to have on average larger absolute brain volumes with respect to females. This result has been confirmed by the analysis of brain volume features obtained both with FS and SPM on the Kirby data sample. By contrast, on the OASIS data sample, this effect is not noticeable, except for one of the subcortical ROIs, consistently for both SPM and FS calculated measures.

The only concerning inconsistency between the two methods we could identify with this case study is the significant greater GM volume of the male cohort, obtained only in the analysis of the FS measures. This kind of inconsistent findings, obtained by straightforward comparisons of the volume measures provided by two widely used segmentation tools, can definitely affect the results of neuroimaging studies and their interpretation and lead to irreproducible results in the literature.

Among the possible limitations of this study is the choice of the selected ROIs. With respect to segmented brain tissues, we limited to GM and WM because the CSF is not provided by FS. The choice of the four subcortical regions (hippocampus, putamen, caudate and brainstem) included both larger and smaller structures, characterized either by different contrasts or by hardly defined borders with respect to surrounding tissues. Another possible limitation is the quite small number of subjects available for each dataset (21 subjects for Kirby and 20 for OASIS). Nevertheless, these data samples allowed identifying systematic intra- and inter-method discrepancies. In addition, this sample size corresponds to the minimal dataset size generally adopted in research studies.

It was not possible to make an absolute evaluation of the accuracy of the segmentation methods, because no gold standard segmentation is available for the data samples we analyzed.

Recently, with the introduction of artificial intelligence, research is heading towards new solutions [40,41] that have the potential to

Table 4
Gender differences in brain structures volumes of interest [ml] for SPM and FS evaluated on the Kirby and OASIS dataset: mean, standard deviation (SD) and statistic measures, t, p-value and Cohen's d. **on p-value > 0.05.*

Kirby dataset										
	FS					SPM				
Brain region	Mean ± SD		Statistic measures			Mean ± SD		Statistic measures		
	Male	Female	t	p-value	Cohen's d	Male	Female	t	p-value	Cohen's d
GM	677 ± 34	590 ± 57	4.0	0.001	1.9	744 ± 32	663 ± 59	3.7	0.002	1.7
WM	543 ± 33	485 ± 53	2.8	0.01	1.3	472 ± 31	423 ± 46	2.7	0.02	1.2
Hippocampus	8.7 ± 0.7	7.9 ± 0.7	2.4	0.02	1.1	7.1 ± 0.3	6.4 ± 0.5	3.5	0.003	1.6
Putamen	11 ± 0.8	9.8 ± 1.1	2.6	0.02	1.2	7.9 ± 0.4	6.8 ± 0.7	4.1	0.001	1.9
Caudate	7.6 ± 0.8	6.7 ± 0.8	2.5	0.02	1.2	6.0 ± 0.4	5.3 ± 0.5	3.3	0.004	1.5
Brainstem	22.4 ± 1.6	18.6 ± 2.7	3.7	0.002	1.7	17 ± 1	14.5 ± 1.8	3.6	0.003	1.7

OASIS dataset										
	FS					SPM				
Brain region	Mean ± SD		Statistic measures			Mean ± SD		Statistic measures		
	Male	Female	t	p-value	Cohen's d	Male	Female	t	p-value	Cohen's d
GM	714 ± 39	662 ± 60	2.2	0.04	1.0	800 ± 42	765 ± 72	1.3	0.21*	0.6
WM	528 ± 35	513 ± 63	0.6	0.53*	0.3	482 ± 34	456 ± 53	1.3	0.22*	0.6
Hippocampus	8.4 ± 0.6	7.8 ± 0.6	1.9	0.08	0.9	7.0 ± 0.3	6.6 ± 0.6	1.9	0.07	0.8
Putamen	11.5 ± 0.9	10.2 ± 1.1	2.8	0.01	1.3	8.0 ± 0.5	7.2 ± 0.9	2.3	0.03	1.0
Caudate	7.9 ± 0.8	7.1 ± 1.1	1.6	0.13*	0.7	6.1 ± 0.5	5.7 ± 0.7	1.5	0.16*	0.6
Brainstem	22.5 ± 2.1	21.3 ± 2.1	1.1	0.28*	0.5	16 ± 1	15.0 ± 1.7	1.5	0.16*	0.6

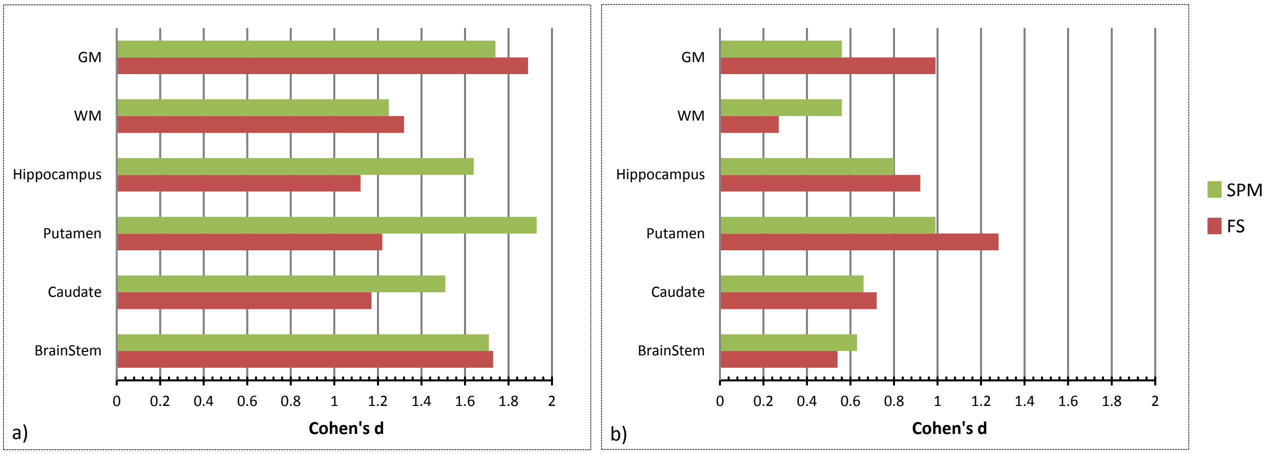

Fig. 12. Histograms of the Cohen's d effect sizes (a) on the Kirby-21 dataset, (b) on the OASIS dataset. Green bars indicate the gender effect size obtained with SPM; red bars indicate the gender effect size obtained with FS.

overcome some limits that these approaches still present. Therefore, it could be useful to apply the presented analysis also to the new machine learning-based segmentation algorithms.

5. Conclusions

This paper has provided a comparison between SPM and FS in terms of the intra-method repeatability and inter-method reproducibility of ROI volumes, evaluated on two different data sets. The considerations above lead us to support SPM as a more consistent tool to evaluate ROI volumes. In any case, as the two methods rely on different algorithm pipelines, which can be differently affected by the presence of abnormalities, image artifacts, or variations in the acquisition protocol parameters, we suggest to cross-validate the findings of each research study against different segmentation methods before proceeding to their interpretation.

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.

Acknowledgments

The OASIS project was funded by grants P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

This work has been partially funded by the Tuscany Government (Bando FAS Salute by Sviluppo Toscana, ARIANNA Project), and by the National Institute of Nuclear Physics (AIM project).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2019.07.016.

References

[1] Hogan RE, Mark KE, Choudhuri I, Wang L, Joshi S, Miller MI, et al. Magnetic resonance imaging deformation-based segmentation and temporal lobe epilepsy. J Digit Imaging 2000;13:217–8. https://doi.org/10.1053/jdim.2000.6897.

[2] Sachdeva J, Kumar V, Gupta I, Khandelwal N, Ahuja CK. Segmentation, feature extraction, and multiclass brain tumor classification. J Digit Imaging 2013;26:1141–50. https://doi.org/10.1007/s10278-013-9600-0.

[3] Akhil M, Aishwarya R, Lal V, Mahesh S. Comparison and evaluation of segmentation techniques for brain mri using Gold Standard. Indian. J Sci Technol 2016;9. https://doi.org/10.17485/ijst/2016/v9i46/106495.

[4] Maclaren J, Han Z, Vos SB, Fischbein N, Bammer R. Reliability of brain volume measurements: a test-retest dataset. Sci Data 2014;1:1–9. https://doi.org/10.1038/sdata.2014.37.

[5] Chard DT, Parker GJM, Griffin CMB, Thompson AJ, Miller DH. The reproducibility and sensitivity of brain tissue volume measurements derived from an SPM-based segmentation methodology. J Magn Reson Imaging 2002;15:259–67. https://doi.org/10.1002/jmri.10064.

[6] Selgrade ES, Wagner HR, Huettel SA, Wang L, McCarthy G, Morey RA, et al. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Hum Brain Mapp 2010;00:1751–62. https://doi.org/10.1002/hbm.20973.

[7] Ochs AL, Ross DE, Zannoni MD, Abildskov TJ, Bigler ED. For the Alzheimer's disease neuroimaging initiative. Comparison of automated brain volume measures obtained with neuroQuant® and FreeSurfer. J Neuroimaging 2015;25:721–7. https://doi.org/10.1111/jon.12229.

[8] Katuwal GJ, Baum SA, Cahill ND, Dougherty CC, Evans E, Evans DW, et al. Inter-method discrepancies in brain volume estimation may drive inconsistent findings in autism. Front Neurosci 2016;10.. https://doi.org/10.3389/fnins.2016.00439.

[9] Wenger E, Mårtensson J, Noack H, Bodammer NC, Kühn S, Schaefer S, Heinze HJ, Düzel E, Bäckman L, Lindenberger ULM. Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains. Hum Brain Mapp 2014;2914:4236–48. https://doi.org/10.1002/hbm.22473.

[10] Kazemi K, Noorizadeh N. Quantitative comparison of SPM, FSL, and brainsuite for brain mr image segmentation. J Biomed Phys Eng 2014;4:13–26.

[11] Perlaki G, Orsi G, Plozer E, Altbacker A, Darnai G, Nagy SA, et al. Are there any gender differences in the hippocampus volume after head-size correction? A volumetric and voxel-based morphometric study. Neurosci Lett 2014;570:119–23. https://doi.org/10.1016/j.neulet.2014.04.013.

[12] Battaglini M, Jenkinson M, De Stefano N. SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI. Hum Brain Mapp 2018. https://doi.org/10.1002/hbm.23828.

[13] Perlaki G, Horvath R, Nagy SA, Bogner P, Doczi T, Janszky J, et al. Comparison of accuracy between FSL's FIRST and Freesurfer for caudate nucleus and putamen segmentation. Sci Rep 2017;7:1–9. https://doi.org/10.1038/s41598-017-02584-5.

[14] Tae WS, Kim SS, Lee KU, Nam EC, Kim KW. Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. Neuroradiology 2008;50:569–81. https://doi.org/10.1007/s00234-008-0383-9.

[15] Jovicich Jorge, Czanner Silvester, Han Xiao, Salat David, van der Kouwe Andre, Quinn Brian, et al. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. Neuroimage 2009;46:177–92.

[16] Barnes J, Ridgway GR, Bartlett J, Henley SMD, Lehmann M, Hobbs N, et al. Head size, age and gender adjustment in MRI studies: a necessary nuisance? Neuroimage 2010;53:1244–55. https://doi.org/10.1016/j.neuroimage.2010.06.025.

[17] Neuroimaging B members & collaborations of the WCFH. SPM, Statistical Parametric Mapping n.d. Available at: https://www.fil.ion.ucl.ac.uk/spm.

[18] Ashburner J, Barnes G, Chen C, Daunizeau J, Moran R, Henson R, et al. SPM12 manual the FIL methods group (and honorary members). Funct Imaging Lab 2013:475–81. https://doi.org/10.1111/j.1365-294X.2006.02813.x.

[19] Imaging L for CNAAMC for BBF. FreeSurfer n.d. https://surfer.nmr.mgh.harvard.edu.

[20] Fischl B. FreeSurfer. Neuroimage 2012. https://doi.org/10.1016/j.neuroimage.2012.01.021.

[21] Landman BA, Huang AJ, Gifford A, Vikram DS, Lim IAL, Farrell JAD, et al. Multi-parametric neuroimaging reproducibility: a 3-T resource study. Neuroimage 2011;54:2854–66. https://doi.org/10.1016/j.neuroimage.2010.11.047.

[22] Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 2007;19:1498–507. https://doi.org/10.1162/jocn.2007.19.9.1498.

[23] NITRC. NeuroImaging Tools & Resources Collaboratory n.d. Available at: https://www.nitrc.org/projects/multimodal.

[24] NITRC. NeuroImaging Tools & Resources Collaboratory. n.d. Available at: https://www.nitrc.org/projects/oasis/.

[25] Klein A, Tourville J. 101 labeled brain images and a consistent human cortical labeling protocol. Front Neurosci 2012;6:1–12. https://doi.org/10.3389/fnins.2012.00171.

[26] Vaz S, Falkmer T, Passmore AE, Parsons R, Andreou P. The case for using the repeatability coefficient when calculating test-retest reliability. PLoS ONE 2013;8:1–7. https://doi.org/10.1371/journal.pone.0073990.

[27] Fischl B, van Der Kouwe A, Salat DH, Busa E, Albert M, Dieterich M, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–55. https://doi.org/10.1016/S0896-6273(02)00569-X.

[28] Morey RA, Petty CM, Xu Y, Pannu Hayes J, Wagner HR, Lewis DV, et al. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 2009;45:855–66. https://doi.org/10.1016/j.neuroimage.2008.12.033.

[29] Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005;26:839–51. https://doi.org/10.1016/j.neuroimage.2005.02.018.

[30] Myles PS, Cui JI. Using the Bland-Altman method to measure agreement with repeated measures. Br J Anaesth 2007;99:309–11. https://doi.org/10.1093/bja/aem214.

[31] Takahashi R, Ishii K, Kakigi T, Yokoyama K. Gender and age differences in normal adult human brain: voxel-based morphometric study. Hum Brain Mapp 2011;32:1050–8. https://doi.org/10.1002/hbm.21088.

[32] Ritchie SJ, Cox SR, Shen X, Lombardo MV, Reus LM, Alloza C, et al. Sex differences in the adult human brain: evidence from 5216 UK biobank participants. Cereb Cortex 2018;28:2959–75. https://doi.org/10.1093/cercor/bhy109.

[33] Ruigrok ANV, Salimi-Khorshidi G, Lai MC, Baron-Cohen S, Lombardo MV, Tait RJ, et al. A meta-analysis of sex differences in human brain structure. Neurosci Biobehav Rev 2014;39:34–50. https://doi.org/10.1016/j.neubiorev.2013.12.004.

[34] Cohen J. Statistical power analysis for the behavioral sciences, second edition. 1988. doi:10.1234/12345678.

[35] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15. https://doi.org/10.1186/s12880-015-0068-x.

[36] Tudorascu DL, Karim HT, Maronge JM, Alhilali L, Fakhran S, Aizenstein HJ, et al. Reproducibility and bias in healthy brain segmentation: comparison of two popular neuroimaging platforms. Front Neurosci 2016;10:1–8. https://doi.org/10.3389/fnins.2016.00503.

[37] Heinen R, Bouvy WH, Mendrik AM, Viergever MA, Biessels GJ, De Bresser J. Robustness of automated methods for brain volume measurements across different MRI field strengths. PLoS ONE 2016. https://doi.org/10.1371/journal.pone.0165719.

[38] Seiger R, Ganger S, Kranz GS, Hahn A, Lanzenberger R. Cortical thickness estimations of FreeSurfer and the CAT12 toolbox in patients with Alzheimer's disease and healthy controls. J Neuroimaging 2018;28:515–23. https://doi.org/10.1111/jon.12521.

[39] Collins DL, Neelin P, Peters TM, Evans AC. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. J Comput Assist Tomogr 1994. https://doi.org/10.1097/00004728-199403000-00005.

[40] Wachinger C, Reuter M, Klein T. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. Neuroimage 2018. https://doi.org/10.1016/j.neuroimage.2017.02.035.

[41] Chen H, Dou Q, Yu L, Qin J, Heng P-A. VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. Neuroimage 2018. https://doi.org/10.1016/j.neuroimage.2017.04.041.