






Article

Evaluating the effect of intensity standardisation on longitudinal whole brain atrophy quantification in brain magnetic resonance imaging

Emily E. Carvajal-Camelo¹, Jose Bernal², Arnau Oliver³, Xavier Lladó³, María Trujillo¹
and The Alzheimer's Disease Neuroimaging Initiative[†]

¹ Multimedia and Computer Vision Group, Universidad del Valle, Cali, Colombia

² Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK

³ Computer Vision and Robotics Institute, Universitat de Girona, Girona, Spain

* Correspondence: emily.carvajal@correounivalle.edu.co and jose.bernal@ed.ac.uk

† Membership list can be found in the Acknowledgments section

Abstract: Atrophy quantification is fundamental for understanding brain development and diagnosing and monitoring brain diseases. FSL-SIENA is a well-known fully-automated method that has been widely used in brain magnetic resonance imaging studies. However, intensity variations arising during image acquisition that may compromise evaluation, analysis and even diagnosis. In this work, we study whether intensity standardisation can improve longitudinal atrophy quantification. We considered seven methods comprising z-score, fuzzy c-means, Gaussian mixture model, kernel density, histogram matching, white stripe, and removal of artificial voxel effects by linear regression (RAVEL). We used a total of 330 scans from two publicly-available datasets, ADNI and OASIS. In scan-rescan assessments, that measures robustness to subtle imaging variations, intensity standardisation did not compromise the robustness of FSL-SIENA significantly ($p > 0.1$). In power analysis assessments, that measures the ability to discern between two groups of subjects, three methods led to consistent improvements in both datasets with respect to the original: fuzzy c-means, Gaussian mixture model, and kernel density estimation. Reduction in sample size using these three methods ranged from 17% to 95%. The performance of the other four methods was affected by spatial normalisation, skull stripping errors, presence of periventricular white matter hyperintensities, or tissue proportion variations over time. Our work evinces the relevance of appropriate intensity standardisation in longitudinal cerebral atrophy assessments using FSL-SIENA.

Keywords: Intensity standardisation; FSL-SIENA; longitudinal atrophy quantification; brain magnetic resonance imaging

1. Introduction

Longitudinal brain atrophy quantification is an active research area in medical image analysis since these measurements permit studying brain development, diagnosing brain diseases and evaluating disease progression over time, and assessing treatment effectiveness [1–6]. Although slow shrinkage of the brain comes with ageing, it may be also a neuroimaging feature of pathologies, such as schizophrenia, cerebral small vessel disease, Alzheimer's disease, and Multiple Sclerosis [7–11]. Therefore, accurate and reliable brain volume measurements are essential for characterising normal and abnormal brain tissue changes and understanding the nature of brain problems.

Longitudinal brain volumetry assessments consist of finding and quantifying brain tissue variations between two scans of the same patient taken at different time-points, a *baseline* and a *follow-up* scan, e.g. scans acquired at inclusion and a year later. Numerous algorithms have been

proposed for carrying out such evaluations [7]. However, FSL-SIENA (Structural Image Evaluation, using Normalization, of Atrophy) [12] continues being a widespread tool in the medical community, given the fact that it is fully automated and available under open source license [7].

Like many other methods, FSL-SIENA computes brain edge displacement as a surrogate measure of atrophy [7]. The processing pipeline comprises skull stripping, registration, tissue segmentation, and brain edge displacement estimation. Naturally, the accuracy and precision of each step determines the overall performance of FSL-SIENA. For example, intensity variations between baseline and follow-up scans – e.g. caused by imaging protocol – may have serious consequences on both, the registration and segmentation steps, affecting subsequent stages of the analysis [13,14].

Previous works have explored ways to improve the FSL-SIENA processing pipeline using other skull stripping algorithms, improving segmentation, or reducing intensity variations [13,15,16]. Shah *et al.* [15] demonstrated the effectiveness of histogram matching for improving multiple sclerosis lesion segmentation in a multi-site multi-scanner setup. Battaglini *et al.* [13] showed the relevance of improved brain extraction and intensity correction modules. In both aforementioned works, the intensity standardisation step consisted of a piece-wise linear histogram matching, which assumes that the balance of tissue classes is consistent between subjects being matched. However, this assumption does not necessarily hold for longitudinal studies [17].

In this work, we evaluate the effect of intensity standardisation strategies on longitudinal atrophy quantification. In particular, we consider seven strategies that are used in medical image analysis, three of them based on traditional clustering algorithms, two which have been recently proposed, two traditional standardisation methods and all the aforementioned are available to the public under open source license [18]. We hypothesise that incorporating intensity standardisation in atrophy quantification assessments leads to significantly better estimations compared to when omitted. To the knowledge of the authors, this is the first time such an analysis has been carried out. The contributions of this work are three-fold: (i) we showcase and make publicly available a ready-to-use tool for assessing the effect of intensity standardisation on scan-rescan and longitudinal atrophy quantification; (ii) we benchmark seven intensity standardisation techniques for harmonising intensities between baseline and follow-up scans within a standard whole brain atrophy quantification pipeline; and (iii) we show quantitatively that intensity standardisation may lead to improved longitudinal atrophy quantification.

2. Materials and methods

2.1. Datasets

We considered two publicly available longitudinal MRI repositories: Open Access Series of Imaging Studies (OASIS) [19] and Alzheimer's Disease Neuroimaging Initiative (ADNI)¹. Both datasets contain MRI scans from with pathological and control subjects. For scan-rescan repeatability, we used the OASIS dataset as it contains scan and rescan data for all patients in the study. Relevant information of each dataset is presented in Table 1. For the sake of reproducibility, we attach the list of selected cases from ADNI and OASIS as supporting documents.

2.2. Equipping FSL-SIENA with intensity standardisation

Our atrophy quantification tool was that of FSL-SIENA [12]. This tool measures brain edge displacement between scans acquired at baseline and follow-up visits as a surrogate measure of cerebral

¹ adni.loni.usc.edu. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Table 1. Relevant information from considered datasets. Although the average reconstruction matrix of the ADNI dataset is the one indicated below, the actual dimensions vary.

Aspect	OASIS2	ADNI2
No. of selected subjects	122	43
Neurodegenerative disease	Dementia	Alzheimer's disease
No. of patients	62	24
No. of control subjects	60	19
Age range	[60 – 96]	[55 – 90]
Field strength	1.5T	3.0T
Voxel spacing	$1.0 \times 1.0 \times 1.3$	$1.2 \times 1.0 \times 1.0$
Reconstruction matrix	$256 \times 256 \times 128$	$196 \times 256 \times 256$
Annual visit	No	Approximately

atrophy. Briefly, the processing pipeline consists of skull stripping, registration, tissue segmentation, and brain edge displacement analysis. In this work, we evaluate the effect of intensity standardisation before and after registration as depicted in Fig. 1, considering that intensity variations may affect both the registration [20] and the segmentation [13].

2.3. Considered intensity standardisation techniques

2.3.1. Z-score

The z-score is a popular normalisation method that linearly transforms intensities to have zero mean and unit standard deviation. Given an input image $I \in \mathbb{R}^{N \times M \times L}$, the process consists of subtracting the mean intensity value from each voxel and dividing the result by the standard deviation value as follows

$$I_{z\text{-score}}(x) = \frac{I(x) - \mu}{\sigma}, \quad (1)$$

where $\mu = \frac{1}{|ICV|} \sum_{v \in ICV} I(v)$, $\sigma = \sqrt{\frac{\sum_{v \in ICV} (I(v) - \mu)^2}{|ICV| - 1}}$, and ICV is the intracranial volume. Since both the mean and standard deviation depend on tissue proportions and data harmonisation quality, this method does not ensure inter-subject standardisation.

2.3.2. Clustering-based white matter mean standardisation

Clustering-based standardisation iteratively classify voxels into regions (for instance, cerebrospinal fluid, grey matter, and white matter) based on intensities values, finding the mean intensity of a reference region (e.g. white matter), and dividing all voxel intensities by such a value [18]. Clustering requires identifying K similar clusters out of the given data points. The selection of K may depend on the task at hand or on the presence of abnormal tissue (e.g. white matter hyperintensities, stroke lesions, tumours) or extra-cerebral regions. In this work, we considered two clustering techniques, fuzzy c-means and Gaussian mixture models, in which they cluster intensities into three groups ($K = 3$, ideally white matter, grey matter, and cerebrospinal fluid), identify the white matter region, and divide all intensities using the mean white matter intensity value [18].

2.3.3. Kernel density estimation

This method estimates a probability density function out of a set of data points (in this case, histogram of intensities). The probability density function, $p(x)$ is expressed as

$$p(x) = \frac{1}{N \cdot M \cdot L} \sum_{x_i \in I} \mathcal{K} \left(\frac{x - x_i}{h} \right), \quad (2)$$

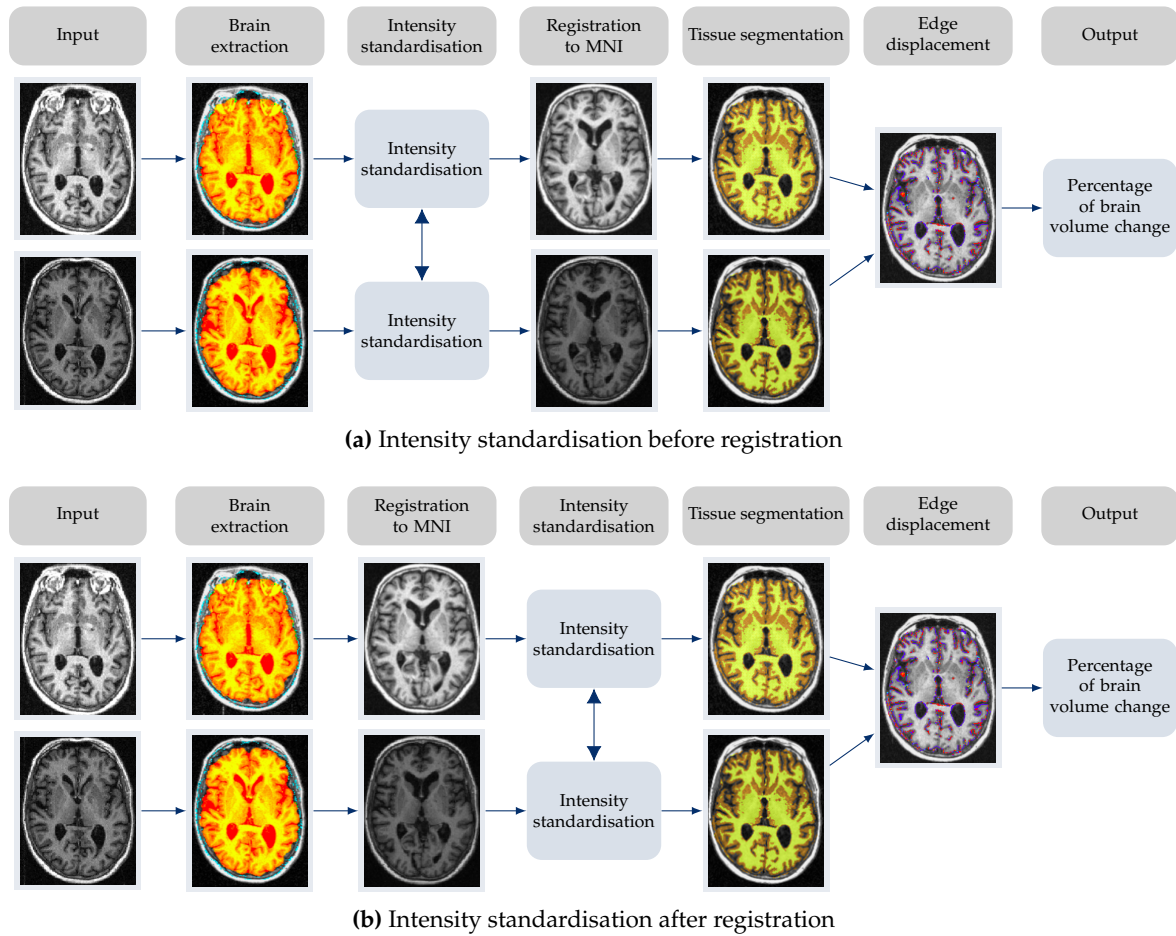


Figure 1. Incorporating intensity standardisation in the FSL-SIENA pipeline. Standardisation takes place before or after registering input volumes to the MNI space. The inputs correspond to the baseline and follow-up T1-w scans. The FSL-SIENA pipeline consists of skull stripping, spatial normalisation, tissue segmentation, edge displacement analysis, and brain volume change reporting.

where \mathcal{K} is the smoothing kernel and h is the smoothing parameter. The contribution of each data point depends on its neighbourhood: the closer the points to x , the higher the $p(x)$. The bandwidth controls the smoothness of $p(x)$: the lower the h , the smoother the resulting probability density function. Once the method computes $p(x)$, it finds its largest peak and assumes it represents the mean intensity value of the normal-appearing white matter. In this implementation, we used a Gaussian kernel and $h = \max(I)/80$ as provided in the implementation in [18].

2.3.4. Piece-wise linear histogram matching

The piece-wise linear histogram matching method proposed by Nyul and Udupa [21] learns an intensity mapping function from a reference scan and uses it to standardise input scans. In the training step, the method identifies a set of p landmarks (e.g. percentiles or modes) from the histogram of the reference scan and creates a piece-wise linear function by interpolating linear segments between consecutive landmarks. The method avoids the minimum and maximum values to add robustness against outliers. In the transformation phase, the same landmarks are identified and intensities updated according to the estimated mapping function.

The histogram matching method assumes equal tissue proportion [17], i.e. histograms of both the reference and input scan are equivalent. However, in longitudinal atrophy studies, tissue proportions vary depending on the tissue loss and brain lesion burden variation between visits. Additionally, differences in imaging or preprocessing quality (skull stripping errors) can lead to an

incorrect identification of landmarks and, hence, an incorrect estimation of the mapping function [22]. Additionally, this approximation presents information loss as direct consequence of the linear interpolation between landmarks.

2.3.5. White stripe

The white stripe method proposed by Shinobara *et al.* [17] estimates the mean and the standard deviation of intensities in the white matter and normalises intensities in a z-score fashion (Eq. 1). The estimation of both parameters is carried out in three steps. First, the method registers the Montreal Neurological Institute (MNI) atlas to the input case to approximately find the white matter. Second, the method finds the largest non-background peak from the histogram of a square region of 4 cm around the centre of the head. The value of the peak is assumed to be the mean intensity of the normal-appearing white matter μ_{WS} . Third, the method computes the standard deviation σ_{WS} from a window of 10% of the intensity values around the mean.

This method assumes that the area around the ventricles contains white matter primarily. Nonetheless, such an assumption may not hold depending on the atrophy extent and overall burden of periventricular white matter hyperintensities. Moreover, the width of the intensity window used for computing the standard deviation was specifically set for a multiple sclerosis cohort. This implies that the method may under- or over-estimate the actual value and, in either case, compromise the standardisation.

2.3.6. Removal of artificial voxel effect by linear regression

The removal of artificial voxel effect by linear regression (RAVEL) method proposed by Fortin *et al.* [23] improves the white stripe standardisation by removing unknown and unwanted technical variability. The method consists of estimating cohort-wise unwanted factors, Z^T , from cerebrospinal fluid regions and subtract it from the intensities obtained after applying the white stripe standardisation method, as described by the following formula

$$I_{RAVEL}^i(x) = I_{WS}^i(x) - \gamma(x) \cdot Z^T \quad (3)$$

where I_{WS}^i represents the intensity obtained after white stripe for the i -th scan and γ_x the weight of the unwanted factors. These unwanted factors and their weights are estimated from the intensities in the cerebrospinal fluid regions through singular value decomposition and linear regression, respectively. Note that the method relies on the standardisation result of white stripe and, hence, white stripe errors propagate to RAVEL.

2.4. Evaluation analysis and measures

2.4.1. Quality of intensity standardisation

The Kullback-Leibler (KL) divergence measures the difference between two probability distributions, expressed as intensity histograms for the task at hand. As the standardisation process seeks to map intensity distributions to a similar range, we use this metric as a measure of the standardisation quality. Given two probability density functions, p and q , the KL divergence is computed as follows

$$KL(p, q) = \sum_{i=1}^N p(x_i) \log \frac{p(x_i)}{q(x_i)}. \quad (4)$$

The more similar the distributions, the lower the KL divergence value. We compared the KL divergence between baseline and follow-up scans before and after standardisation using the Wilcoxon signed-rank test. We expect the KL divergence value to decrease after standardising intensities.

2.4.2. Scan-rescan repeatability

The scan-rescan error measures the robustness of an atrophy quantification algorithm against subtle imaging variations. Since the atrophy level between scans of the same patient taken on the same visit with the same imaging protocol and same scanner should be minimal, the expected percentage of brain volume change measured by FSL-SIENA should be close to zero. We studied whether standardising intensities could affect the performance of the original method. We compared the scan-rescan deviation with and without equipping FSL-SIENA with the intensity standardisation step using the Wilcoxon signed-rank test.

2.4.3. Power analysis for longitudinal atrophy quantification

The standard comparison of longitudinal atrophy quantification methods consists of examining their ability to discern between two groups of subjects undergoing different treatments/pathologies in terms of atrophy measurements (e.g. OASIS: dementia patients vs control subjects; ADNI: Alzheimer's disease patients vs control subjects). This evaluation requires MRI intervals to be consistent across subjects. Although in ADNI, this is approximately the case, it is not in OASIS. Thus, we annualised volume change values as in [16]. Due to possible outliers in the data, we considered the following formulation of the effect size [24,25]:

$$\text{Effect size} = 2\sqrt{\frac{\eta^2}{1 - \eta^2}}, \quad (5)$$

where $\eta^2 = \frac{Z^2}{N-1}$ and Z is the standardised value for the U-value of the Mann-Whitney U-test. Of note, the higher the effect size, the better the algorithm at discerning between groups. With this information, we calculated the required sample size, i.e. the minimum number of cases in each group to detect such an effect, as follows:

$$\text{Sample size} = \left\lceil 2 \left(\frac{u + v}{\text{TxE} \cdot \text{Effect size}} \right)^2 \right\rceil, \quad (6)$$

where $u = 0.842$ for 80% power, $v = 1.96$ to test at 5% significance level, $\lceil \cdot \rceil$ is the ceiling operator, and $\text{TxE} = 25\%$ is the predefined treatment effect. Of note, this effect is assumed to be immediate and constant throughout the evaluation interval. We included all percentages of brain volume change values yielded by FSL-SIENA in the analysis.

2.5. Implementation details

We implemented all methods in Python using the Intensity Normalisation library [18] and used FSL-SIENA v6.0. We ran all the experiments on a GNU/Linux machine box running Ubuntu 18.04, with 16 GB RAM. The developed framework is available to download at our Github repository². We carried out all statistical analyses using RStudio v1.1.456 with R v3.5.1.

3. Results

We ran FSL-SIENA on both ADNI and OASIS before and after standardisation with the proposed pipelines shown in Fig. 1 and assessed the effect of the additional processing step on its repeatability and statistical power. Also, we evaluated the similarity between histogram of intensities of baseline and follow-up scans before and after standardisation using the KL divergence as a surrogate measure of the standardisation quality. The experimental results are described in the following sections and included as supporting documents.

² Docker containing the implementation can be found at <https://github.com/emysme/IntensityStandardisation>

Table 2. Sample size obtained using FSL-SIENA using different skull stripping approximations. We assumed 80% power, 25% treatment effect, and 5% significance level. IQR: Interquartile range.

	Skull stripping technique				
	BET	BET "-f 0.1"	BET "-f 0.2"	BEaST	ROBEX
OASIS (n=123)					
Median (IQR) Demented, %	-1.214 (-3.010, -0.412)	-1.913 (-3.160, -0.445)	-1.963 (-2.500, -0.987)	-1.800 (-2.822, -0.891)	-0.463 (-0.820, -0.292)
Median (IQR) Control, %	-0.813 (-1.590, -0.216)	-0.592 (-1.210, 0.144)	-0.652 (-1.160, -0.207)	-0.781 (-1.276, -0.379)	-0.159 (-0.264, -0.043)
Effect size	0.703	0.803	1.042	0.983	1.141
Sample size, n	509	390	232	205	152
Sample size improvement, %	-	23.379	54.420	59.724	70.138

3.1. Ablation experiment

Given that skull stripping errors have a significant detrimental effect on the performance of FSL-SIENA, as demonstrated in previous works [13,16], we first determined which skull stripping strategy yielded the lowest sample size. We used FSL-SIENA [12] with skull stripping approaches, BET [26] (default), BEaST (used in Nakamura et al. [16]) and ROBEX [27]. The results are condensed in Table 2. Skull stripping with ROBEX led to the lowest sample size and highest improvement compared to the baseline (70%). Based on this outcome, we opted for removing extra-cerebral regions using ROBEX.

3.2. Quality of intensity standardisation

We measured the quality of standardisation by determining whether histograms of baseline and follow-up scans were similar, in terms of the KL divergence, after standardising intensities or not. The results are presented in Fig. 2. In most cases, standardising intensities between baseline and follow-up scans resulted in lower KL divergence values. The divergence values were even lower when standardising intensities before registering scans to the MNI space. In particular, the two cluster-based strategies (fuzzy c-means and Gaussian mixture models) obtained the lowest KL divergence values compared to the rest of the methods. On the other hand, the application of white stripe and RAVEL was not beneficial as KL divergence values tended to be higher, i.e. higher mismatch between histograms of intensities, compared to those obtained when scans were not processed at all.

3.3. Scan-rescan repeatability

The scan-rescan repeatability experiment consisted of examining whether intensity standardisation could result in significant differences – for better or worse – in scan-rescan assessments against the baseline. We ran FSL-SIENA on pairs of scans of $n = 122$ subjects that were acquired in a single session with the same protocol and same scanner before and after standardisation (OASIS dataset). We used the resulting percentage of brain volume change yielded by FSL-SIENA as measure of error since this variation should be close to zero. The results are presented in Fig. 3. None of the considered intensity standardisation methods compromised the baseline robustness of FSL-SIENA against subtle imaging artefacts ($p > 0.1$). However, the application of z-score, white stripe and RAVEL before registration increased the variance of atrophy measurements.

3.4. Power analysis for longitudinal atrophy quantification

Power analysis for longitudinal atrophy quantification consisted of determining whether intensity standardisation could enhance the ability of FSL-SIENA to discern two groups of subjects undergoing different pathologies (OASIS: demented vs control; ADNI: Alzheimer's disease vs control). For that, we examined the effect size. The results are presented in Table 3. In OASIS, standardising intensities with the clustering-based or kernel density estimation approximations led to a reduction in the sample size between 18% and 23%. The application of other intensity standardisation techniques resulted in lower effect sizes or varied notably depending whether they were applied before or after spatial normalisation. In ADNI, sample size improvements were much higher than in OASIS, reaching

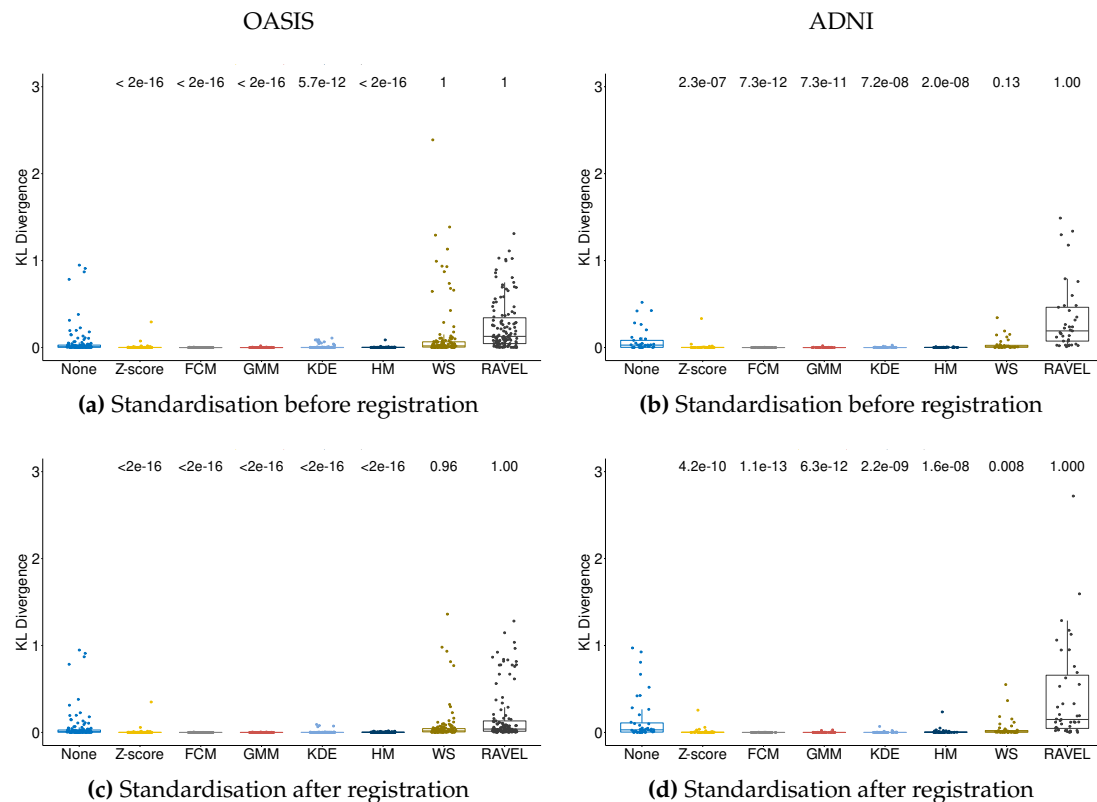


Figure 2. Similarity between histograms of intensities with and without standardisation using the considered strategies on OASIS and ADNI. We equipped FSL-SIENA with intensity standardisation before or after registration. We measured the degree of similarity between the histogram of intensities of baseline and follow-up using the Kullback-Leibler (KL) divergence. FCM: Fuzzy c-means. GMM: Gaussian mixture model. KDE: Kernel density estimation. HM: Histogram matching. WS: White stripe. RAVEL: Removal of artificial voxel effect by linear regression.

reductions in sample size of up to 95% compared to that obtained using the original FSL-SIENA. The use of most intensity standardisation techniques, except for white stripe and RAVEL, helped FSL-SIENA to discern better between Alzheimer's disease patients and control subjects. Applying intensity standardisation before spatial normalisation resulted in higher effect sizes in this dataset. In both datasets, the absolute atrophy values reported for patients with neurodegenerative diseases were higher than those of the control group.

4. Discussion

Recent works have shown that intensity non-standardness leads to atrophy quantification errors [13]. In this work, we studied whether intensity standardisation could improve longitudinal whole brain atrophy quantification. In particular, we examined the effect of such a harmonisation strategy on an established and thoroughly validated tool provided in the FSL package, FSL-SIENA. We considered seven intensity standardisation techniques comprising z-score, fuzzy c-means, Gaussian mixture model, kernel density estimation, histogram matching, white stripe, and removal of artificial voxel effects by linear regression (RAVEL). We evaluated their effect on the scan-rescan repeatability and statistical power of FSL-SIENA. We used 165 pairs of baseline and follow-up scans from patients with neurodegenerative diseases and control subjects from ADNI and OASIS (122 and 43, respectively). To our knowledge, this is the first time that the effect and suitability of multiple intensity standardisation has been quantitatively investigated for cerebral atrophy quantification.

The aim of intensity standardisation is to reduce intensity variations arising during imaging. Except for white stripe and RAVEL, the considered intensity standardisation methods indeed reduced

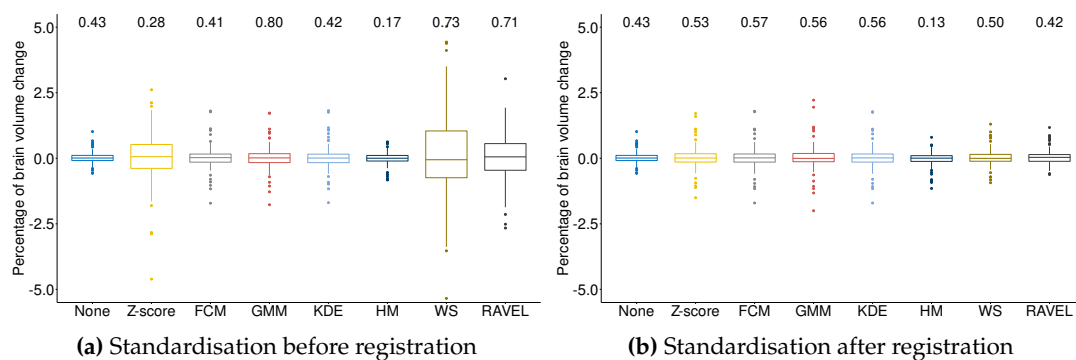


Figure 3. Scan-rescan deviation with and without intensity standardisation. We equipped FSL-SIENA with intensity standardisation before or after registration. We expressed the error as the percentage of brain volume change between scans acquired during the same visit and using the same scanner and acquisition protocol. We used the Wilcoxon signed-rank test to examine whether the median error obtained after standardisation differed from that of the baseline. FCM: Fuzzy c-means. GMM: Gaussian mixture model. KDE: Kernel density estimation. HM: Histogram matching. WS: White stripe. RAVEL: Removal of artificial voxel effect by linear regression.

the original inter-scan intensity variability significantly ($p < 0.001$). Equipping FSL-SIENA with fuzzy c-means, Gaussian mixture model and kernel density based standardisation schemes led to no significant changes in scan-rescan assessments ($p > 0.1$) and improved power (sample size reductions between 17% to 95%). Given that the performance of clustering-based methods depends on their ability to find K clusters in the data, we recommend incorporating the intensity standardisation scheme based on the detection of the white matter mean value using kernel density estimation into the FSL-SIENA pipeline. The effect of the other four intensity standardisation methods varied depending on whether spatial normalisation happened before or after standardising intensities, skull stripping errors, and tissue variations over time. The use of intensity standardisation before registration in both datasets showed a larger improvement (from 18% to 23% in OASIS and from 77% to 83% in ADNI with kernel density estimation).

The white stripe method has assumptions that may not be fulfilled in longitudinal studies. First, the presence of the normal-appearing white matter is conditioned by the presence of periventricular white matter hyperintensities. The higher the burden of hyperintensities, the less the normal tissue. Second, the method computes the standard deviation within a window of interest. This decision may result in an over- or under- estimation of the actual value and thus compromise the standardisation process. Naturally, these problems present in the white stripe method are inherited by RAVEL.

The histogram matching method does not work consistently well for longitudinal studies since tissue proportions vary over time and also does not preserve information as it performs linear interpolations between landmarks. On the one hand, the higher the number of landmarks, the more the method preserves the information, but the more it assumes similar tissue proportion between input scans. On the other hand, the lower the number of landmarks, the higher the information loss. A compromise between these two factors should be reached for the method to be considered in longitudinal assessments.

Future work should contemplate three aspects. First, even though FSL-SIENA is highly accessible and relatively easy and ready to use, we are aware that it is not in the state-of-the-art of longitudinal atrophy quantification in multiple sclerosis and Alzheimer's disease, but the Jacobian determinant integration method [28,29]. Nonetheless, intensity variations may compromise atrophy quantification as well as other types of analyses. Thus, future work should consider extending this study to other techniques. Second, the techniques we examined in this work dispense with the spatial information and, thus, they may be affected by noise and intensity inhomogeneities. Hence, future work should

Table 3. Sample size using FSL-SIENA with and without standardisation. We equipped FSL-SIENA with intensity standardisation before or after registration. We assumed 80% power, 25% treatment effect, and 5% significance level. More details are given in Table A1. Asterisks indicate statistics were computed using less data as FSL-SIENA failed to process scans standardised with RAVEL. RAVEL: Removal of artificial voxel effect by linear regression.

		None	z-score	Fuzzy c-means	Gaussian mixture model	Kernel density estimation	Histogram matching	White stripe	RAVEL
OASIS (n=122)	before	Effect size	1.141	1.02	1.287	1.285	1.301	0.822	0.532
		Sample size, n	152	191	120	120	117	293	700
		Improvement, %	-	-	21.052	21.052	23.026	-	-
	after	Effect size	1.141	1.243	1.268	1.256	1.268	0.699	1.177
		Sample size, n	152	129	124	126	124	405	155
		Improvement, %	-	15.131	18.421	17.105	18.421	-	5.921
ADNI (n=43)	before	Effect size	0.135	0.677	0.309	0.364	0.333	0.435	0.022
		Sample size, n	10856	432	2073	1494	1785	1046	408765
		Improvement, %	-	96.021	80.905	86.238	83.557	90.365	-
	after	Effect size	0.135	0.309	0.286	0.591	0.286	0.419	0.195
		Sample size, n	10856	2073	2419	567	2419	1127	5203
		Improvement, %	-	80.904	77.717	94.777	77.717	89.619	52.073

evaluate techniques making use of both intensity and spatial information. Third, the domain shift problem continues being a key challenge in machine learning as distributions from training and testing may vary depending on scanners and protocols [30,31]. Intensity harmonisation may reduce variations between training and test set, leading to improved performance.

In conclusion, intensity standardisation can improve longitudinal whole-brain atrophy quantification using FSL-SIENA, but not all intensity standardisation methods do. Their applicability depends to a great extent on whether their theoretical assumptions are met or not. Clustering and kernel density estimation based intensity standardisation methods tend to produce the best results compared to other considered techniques. We recommend using the kernel density estimation technique since it is simpler to compute than clustering.

Author Contributions: Conceptualization, E.E.C.C and J.B.; methodology, E.E.C.C. and J.B.; software, E.E.C.C and J.B.; validation, J.B.; formal analysis, E.E.C.C and J.B.; investigation, E.E.C.C. and J.B.; resources, M.T.; data curation, E.E.C.C.; writing—original draft preparation, E.E.C.C. and J.B.; writing—review and editing, E.E.C.C., J.B., A.O., X.L. and M.T.; visualization, E.E.C.C., J.B., A.O., and X.L.; supervision, J.B and M.T.; project administration, J.B.; funding acquisition, M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the MRC Doctoral Training Programme in Precision Medicine (JB).

Acknowledgments: This work has been partially supported by DPI2017-86696-R from the Ministerio de Ciencia, Innovación y Universidades. Data used in the preparation of this article were [in part] obtained from the OASIS dataset: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. Data collection and sharing for the ADNI project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012)³. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical

Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ADNI	Alzheimer's Disease Neuroimaging Initiative
FCM	Fuzzy c-means
FSL-SIENA	Structural Image Evaluation, using Normalization, of Atrophy
GMM	Gaussian mixture model
HM	Histogram matching
ICV	Intracranial volume
KDE	Kernel density estimation
KL	Kullback-Leibler
MNI	Montreal Neurological Institute
MRI	Magnetic resonance imaging
OASIS	Open Access Series of Imaging Studies
RAVEL	Removal of artificial voxel effects by linear regression
WS	White stripe

Table A1. Minimum number of subjects required to detect atrophy differences between patients and control subjects using the FSL-SIENA with and without intensity standardisation. We equipped FSL-SIENA with intensity standardisation before or after registration. We computed sample size using data from the OASIS dataset (n=122) and ADNI (n=43), assuming 80% power, 25% treatment effect, and 5% significance level. FSL-SIENA could not operate on six ADNI scans standardised using RAVEL before registration. IQR: Interquartile range. RAVEL: Removal of artificial voxel effect by linear regression.

	None	Z-score	Fuzzy c-means	Gaussian mixture model	Kernel density estimation	Histogram matching	White stripe	RAVEL
OASIS								
Standardisation before registration								
Median (IQR) Demented, %	-0.463 (-0.820, -0.292)	-0.890 (-1.807, -0.241)	-0.874 (-1.295, -0.456)	-0.849 (-1.310, -0.456)	-0.880 (-1.317, -0.456)	-0.413 (-0.744, -0.198)	-1.065 (-4.198, 1.022)	-0.314 (-0.766, -0.055)
Median (IQR) Control, %	-0.159 (-0.264, -0.043)	-0.138 (-0.591, 0.390)	-0.197 (-0.417, 0.039)	-0.207 (-0.406, -0.048)	-0.207 (-0.400, -0.033)	-0.212 (-0.294, -0.065)	-0.663 (2.564, 1.015)	-0.111 (-0.364, 0.116)
Effect size	1.141	1.02	1.287	1.285	1.301	0.822	0.198	0.532
Sample size, n	152	191	120	120	117	293	5047	700
Sample size improvement, %	-	-	21.052	21.052	23.026	-	-	-
Standardisation after registration								
Median (IQR) Demented, %	-0.463 (-0.820, -0.292)	-0.807 (-1.361, -0.433)	-0.886 (-1.338, -0.429)	-0.962 (-1.443, -0.508)	-0.882 (-1.339, -0.429)	-0.392 (-0.770, -0.226)	-0.686 (-1.155, -0.226)	-0.556 (-1.042, -0.315)
Median (IQR) Control, %	-0.159 (-0.264, -0.043)	-0.202 (-0.405, -0.063)	-0.225 (-0.439, -0.042)	-0.218 (-0.382, -0.060)	-0.227 (-0.439, -0.043)	-0.197 (-0.344, -0.0431)	-0.179 (-0.355, -0.071)	-1.741 (-0.348, 0.069)
Effect size	1.141	1.243	1.268	1.256	1.268	0.699	1.177	1.132
Sample size, n	152	129	124	126	124	405	143	155
Sample size improvement, %	-	15.131	18.421	17.105	18.421	-	5.921	-
ADNI								
Standardisation before registration								
Median (IQR) Alzheimer's disease, %	-0.467 (-0.909, -0.318)	-2.443 (-3.505, -0.863)	-1.044 (-2.110, -0.475)	-1.110 (-2.092, -0.455)	-1.099 (-2.096, -0.484)	-0.551 (-0.810, -0.302)	-0.235 (-1.547, 0.827)	-0.411 (-1.003, 0.078)
Median (IQR) Control, %	-0.679 (-1.272, -0.759)	0.035 (-2.303, 1.533)	-1.163 (-1.804, 1.233)	-1.082 (-1.679, 1.256)	-1.069 (-1.769, 1.207)	-0.228 (-0.899, 0.407)	-0.087 (-2.131, 1.420)	-1.146 (-1.787)
Effect size	0.135	0.677	0.309	0.364	0.333	0.435	0.022	0.804*
Sample size, n	10856	432	2073	1494	1785	1046	408765	307*
Sample size improvement, %	-	96.021	80.905	86.238	83.557	90.365	-	97.172*
Standardisation after registration								
Median (IQR) Alzheimer's disease, %	-0.467 (-0.909, -0.318)	-1.109 (-1.969, -0.493)	-1.117 (-2.124, -0.469)	-1.253 (-2.390, -0.477)	-1.116 (-1.229, -0.468)	-0.509 (-0.747, -0.268)	-0.739 (-1.284, -0.469)	-0.637 (-1.073, -0.126)
Median (IQR) Control, %	-0.679 (-1.272, -0.759)	-1.091 (-1.692, 1.167)	-1.143 (-1.746, 1.234)	-0.317 (-1.726, 1.293)	-1.145 (-1.746, 1.233)	-0.164 (-0.709, 0.349)	-0.921 (-1.144, 0.840)	-0.645 (-1.454, 0.659)
Effect size	0.135	0.309	0.286	0.591	0.286	0.419	0.195	0.180
Sample size, n	10856	2073	2419	567	2419	1127	5203	6107
Sample size improvement, %	-	80.904	77.717	94.777	77.717	89.619	52.073	43.745

References

1. Rovira, À.; Wattjes, M.P.; Tintoré, M.; Tur, C.; Yousry, T.A.; Sormani, M.P.; De Stefano, N.; Filippi, M.; Auger, C.; Rocca, M.A.; others. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis - clinical implementation in the diagnostic process. *Nature Reviews Neurology* **2015**, *11*, 471–482.
2. Steenwijk, M.D.; Geurts, J.J.G.; Daams, M.; Tijms, B.M.; Wink, A.M.; Balk, L.J.; Tewarie, P.K.; Uitdehaag, B.M.J.; Barkhof, F.; Vrenken, H.; others. Cortical atrophy patterns in multiple sclerosis are non-random and clinically relevant. *Brain* **2016**, *139*, 115–126.
3. Filippi, M.; Rocca, M.A.; Ciccarelli, O.; De Stefano, N.; Evangelou, N.; Kappos, L.; Rovira, A.; Sastre-Garriga, J.; Tintoré, M.; Frederiksen, J.L.; Gasperini, C.; Palace, J.; Reich, D.S.; Banwell, B.; Montalban, X.; Barkhof, F. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *The Lancet Neurology* **2016**, *15*, 292 – 303.
4. Storelli, L.; Rocca, M.A.; Pagani, E.; Van Hecke, W.; Horsfield, M.A.; De Stefano, N.; Rovira, A.; Sastre-Garriga, J.; Palace, J.; Sima, D.; others. Measurement of Whole-Brain and Gray Matter Atrophy in Multiple Sclerosis: Assessment with MR Imaging. *Radiology* **2018**, p. 172468.
5. Kushibar, K.; Valverde, S.; González-Villà, S.; Bernal, J.; Cabezas, M.; Oliver, A.; Lladó, X. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Medical Image Analysis* **2018**, *48*, 177–186.
6. Bernal, J.; Kushibar, K.; Cabezas, M.; Valverde, S.; Oliver, A.; Lladó, X. Quantitative Analysis of Patch-Based Fully Convolutional Neural Networks for Tissue Segmentation on Brain Magnetic Resonance Imaging. *IEEE Access* **2019**, *7*, 89986–90002.
7. Cover, K.S.; van Schijndel, R.A.; van Dijk, B.W.; Redolfi, A.; Knol, D.L.; Frisoni, G.B.; Barkhof, F.; Vrenken, H.; Initiative, A.D.N.; others. Assessing the reproducibility of the SIENAX and SIENA brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. *Psychiatry Research: Neuroimaging* **2011**, *193*, 182–190.
8. Haijma, S.V.; Van Haren, N.; Cahn, W.; Koolschijn, P.C.M.; Hulshoff Pol, H.E.; Kahn, R.S. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia bulletin* **2012**, *39*, 1129–1138.
9. Wardlaw, J.M.; Smith, C.; Dichgans, M. Mechanisms of sporadic cerebral small vessel disease: insights from neuroimaging. *The Lancet Neurology* **2013**, *12*, 483–497.
10. van Erp, T.G.; Hibar, D.P.; Rasmussen, J.M.; Glahn, D.C.; Pearlson, G.D.; Andreassen, O.A.; Agartz, I.; Westlye, L.T.; Haukvik, U.K.; Dale, A.M.; others. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular psychiatry* **2016**, *21*, 547.
1. Rocca, M.A.; Battaglini, M.; Benedict, R.H.; De Stefano, N.; Geurts, J.J.; Henry, R.G.; Horsfield, M.A.; Jenkinson, M.; Pagani, E.; Filippi, M. Brain MRI atrophy quantification in MS: from methods to clinical application. *Neurology* **2016**, pp. 10–1212.
2. Smith, S.M.; Zhang, Y.; Jenkinson, M.; Chen, J.; Matthews, P.; Federico, A.; De Stefano, N. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage* **2002**, *17*, 479–489.
3. Battaglini, M.; Jenkinson, M.; De Stefano, N.; Initiative, A.D.N. SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI. *Human brain mapping* **2018**, *39*, 1063–1077.
4. Lee, H.; Nakamura, K.; Narayanan, S.; Brown, R.A.; Arnold, D.L.; Initiative, A.D.N.; others. Estimating and accounting for the effect of MRI scanner changes on longitudinal whole-brain volume change measurements. *NeuroImage* **2019**, *184*, 555–565.
5. Shah, M.; Xiao, Y.; Subbanna, N.; Francis, S.; Arnold, D.L.; Collins, D.L.; Arbel, T. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis* **2011**, *15*, 267–282.
6. Nakamura, K.; Eskildsen, S.F.; Narayanan, S.; Arnold, D.L.; Collins, D.L.; Initiative, A.D.N.; others. Improving the SIENA performance using BEaST brain extraction. *PloS one* **2018**, *13*, e0196945.
7. Shinohara, R.T.; Sweeney, E.M.; Goldsmith, J.; Shiee, N.; Mateen, F.J.; Calabresi, P.A.; Jarso, S.; Pham, D.L.; Reich, D.S.; Crainiceanu, C.M.; others. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **2014**, *6*, 9–19.

18. Reinhold, J.C.; Dewey, B.E.; Carass, A.; Prince, J.L. Evaluating the impact of intensity normalization on MR image synthesis. *Medical Imaging 2019: Image Processing*. International Society for Optics and Photonics, 2019, Vol. 10949, p. 109493H.
19. Marcus, D.S.; Fotenos, A.F.; Csernansky, J.G.; Morris, J.C.; Buckner, R.L. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of cognitive neuroscience* **2010**, *22*, 2677–2684.
20. Bağcı, U.; Udupa, J.K.; Bai, L. The role of intensity standardization in medical image registration. *Pattern Recognition Letters* **2010**, *31*, 315–323.
21. Nyúl, L.G.; Udupa, J.K.; Zhang, X. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging* **2000**, *19*, 143–150.
22. Jäger, F. *Normalization of magnetic resonance images and its application to the diagnosis of the scoliotic spine*; Vol. 34, Logos Verlag Berlin GmbH, 2011.
23. Fortin, J.P.; Sweeney, E.M.; Muschelli, J.; Crainiceanu, C.M.; Shinohara, R.T.; Initiative, A.D.N.; others. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage* **2016**, *132*, 198–212.
24. Cohen, J. *Statistical power analysis for the behavioral sciences*; Academic press, 2013.
25. Tomczak, M.; Tomczak, E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *TRENDS in Sport Sciences* **2014**.
26. Smith, S.M. Fast robust automated brain extraction. *Human brain mapping* **2002**, *17*, 143–155.
27. Iglesias, J.E.; Liu, C.Y.; Thompson, P.M.; Tu, Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging* **2011**, *30*, 1617–1634.
28. Boyes, R.G.; Rueckert, D.; Aljabar, P.; Whitwell, J.; Schott, J.M.; Hill, D.L.; Fox, N.C. Cerebral atrophy measurements using Jacobian integration: comparison with the boundary shift integral. *NeuroImage* **2006**, *32*, 159–169.
29. Nakamura, K.; Guizard, N.; Fonov, V.S.; Narayanan, S.; Collins, D.L.; Arnold, D.L. Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. *NeuroImage: Clinical* **2014**, *4*, 10–17.
30. Kushibar, K.; Valverde, S.; González-Villà, S.; Bernal, J.; Cabezas, M.; Oliver, A.; Lladó, X. Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Scientific Reports* **2019**, *9*, 1–15.
31. Bernal, J.; Kushibar, K.; Asfaw, D.S.; Valverde, S.; Oliver, A.; Martí, R.; Lladó, X. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine* **2019**, *95*, 64–81.