NeuroImage 223 (2020) 117248

Contents lists available at ScienceDirect

NeuroImage

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

A comparison of Freesurfer and multi-atlas MUSE for brain anatomy segmentation: Findings about size and age bias, and inter-scanner stability in multi-site aging studies

Dhivya Srinivasan[a,1,*], Guray Erus[a,1], Jimit Doshi[a], David A. Wolk[b], Haochang Shou[a,c], Mohamad Habes[a,b,2], Christos Davatzikos[a,2], for the Alzheimer's Disease Neuroimaging Initiative[3]

[a] Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Richards Building, 3700 Hamilton Walk, 7th Floor, Philadelphia, PA 19104, United States

[b] Department of Neurology, University of Pennsylvania, United States

[c] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, United States

ARTICLE INFO

Keywords:
MRI
Segmentation
Freesurfer
MUSE
Brain
ROI

ABSTRACT

Automatic segmentation of brain anatomy has been a key processing step in quantitative neuroimaging analyses. An extensive body of literature has relied on Freesurfer segmentations. Yet, in recent years, the multi-atlas segmentation framework has consistently obtained results with superior accuracy in various evaluations. We compared brain anatomy segmentations from Freesurfer, which uses a single probabilistic atlas strategy, against segmentations from Multi-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters and locally optimal atlas selection (MUSE), one of the leading ensemble-based methods that calculates a consensus segmentation through fusion of anatomical labels from multiple atlases and registrations. The focus of our evaluation was twofold. First, using manual ground-truth hippocampus segmentations, we found that Freesurfer segmentations showed a bias towards over-segmentation of larger hippocampi, and under-segmentation in older age. This bias was more pronounced in Freesurfer-v5.3, which has been used in multiple previous studies of aging, while the effect was mitigated in more recent Freesurfer-v6.0, albeit still present. Second, we evaluated inter-scanner segmentation stability using same day scan pairs from ADNI acquired on 1.5T and 3T scanners. We also found that MUSE obtains more consistent segmentations across scanners compared to Freesurfer, particularly in the deep structures.

1. Introduction

Segmentation of brain anatomy has been a key image processing step in neuroimaging studies, as it enables assessment of regional brain volumes in a range of neurological diseases and conditions for image-based diagnosis, monitoring of disease progression, and tracking of neuro-developmental and aging-related brain changes (Giorgio 2013; Janowitz et al., 2014; Raz et al., 2010; Wierenga et al., 2014). Volumetric analyses of cortical structures provided markers of neurodegeneration in various disorders including multiple sclerosis (MS), schizophrenia (SCZ) and Alzheimer's disease (AD), as well as in normal aging (Bonilha et al., 2008; Brewer et al., 2009; Bakkour et al., 2013; Charil et al., 2007; Dicks et al., 2019). Previous studies also reported significant associations between volumes of sub-cortical deep brain structures, including the thalamus, caudate, putamen and amygdala, and neuropsychiatric and neuro-degenerative conditions such as AD and SCZ, suggesting that both cortical and subcortical structures are variably related to different neurodegenerative conditions (Ferreira et al., 2017; Janowitz et al., 2014; Apostolova et al., 2006; Goldstein et al., 1999; Satterthwaite et al., 2016).

* Corresponding author.
E-mail address: Dhivya.Srinivasan@pennmedicine.upenn.edu (D. Srinivasan).

[1] Denotes equally contributing first authors.

[2] Denotes equally contributing senior authors.

[3] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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.

https://doi.org/10.1016/j.neuroimage.2020.117248
Received 22 May 2020; Accepted 4 August 2020
Available online 27 August 2020

Table 1
General Characteristics of datasets that were used in validation experiments.

	CN	MCI	AD	Total
Dataset1: EADC-ADNI				
Count	44	45	45	134
Age	76.18 ± 7.45	74.44 ± 8.00	74.45 ± 8.10	75 ± 7.84
Sex(M/F)	22/22	26/19	21/24	69/65
Dataset2: ADNI 1.5T, ADNI 3T[*]				
Count	37	53	23	113
Age	75.72 ± 4.20	76.02 ± 7.90	75.03 ± 8.36	75.72 ± 6.97
Sex(M/F)	15/22	35/18	08/15	58/55

[*] For each subject Dataset2 includes a pair of 1.5T and 3T scans acquired on same day.

Multiple algorithms and methods have been developed for segmenting anatomical regions of interest (ROIs) in a fully automated way. One of the most widely used automated segmentation methods is the publicly available FreeSurfer software package (Fischl et al., 2002; 2004). Freesurfer combines a surface-based stream for cortical segmentation with a volume-based stream for segmentation of subcortical structures. The surface-based stream first calculates an initial surface that delineates the white matter, and then refines this surface to calculate the pial surface. The volume-based stream uses a subject-independent probabilistic atlas, which is automatically derived from a training set consisting of multiple hand-labeled atlas images. After a high dimensional nonlinear volumetric alignment of the target image to a common atlas space, ROI labels are automatically assigned to each voxel by finding a segmentation that maximizes the probability of the input data given the prior probabilities from the training set. At present, an important body of literature and neuropsychiatry findings are based on Freesurfer segmentations (Kikinis et al., 2010; Rohrer2011; Messina et al., 2011; Sabuncu et al., 2011). Freesurfer has been widely tested for its accuracy, precision and repeatability. However, various evaluations have also revealed limitations of that approach (Mccarthy 2015; Keller et al., 2012; Mulder et al., 2014), which might be critical in aging and multi-site studies.

Instead of relying on a single atlas, the multi-atlas segmentation (MAS) framework utilizes multiple reference atlas images that are independently warped to the target image space, and their reference labels are fused together to derive a consensus segmentation. This process has important advantages, the most notably the robustness against individual registration errors by the virtue of the ensemble label fusion process. The MUSE algorithm (Doshi et al., 2016) extended the ensemble approach to multiple deformable registration algorithms applied at different regularizations, allowing a higher variation within the ensemble. Additionally, in the label fusion step, MUSE uses a spatially adaptive strategy to select atlases based on their local similarity to the target image. Effectively, atlases most similar to a target image being segmented have more influence on its segmentation. Even more importantly, this process is spatially-adaptive, i.e. atlases most suitable for someone's hippocampus segmentation might not necessarily be best for thalamus segmentation. MUSE was the top-ranking method in the MICCAI-SATA challenge on deep brain segmentation (Asman et al., 2013), and has been used in several studies (Habes et al., 2016; Satterthwaite et al., 2016; Wee 2017; Tian et al., 2018).

Our primary motivation herein was to evaluate brain anatomy segmentations obtained using these two automated methods in multi-site brain aging studies. Importantly, our comparative evaluations included both versions v6.0 and v5.3 of Freesurfer, thus also providing a comparison between newer and older Freesurfer versions. This comparison may be informative for guiding processing pipeline updates in longitudinal studies that previously used older versions of Freesurfer for segmentation. Accurately measuring subtle brain aging changes, and particularly hippocampal volume change, is important for early identification of pathologic processes. Moreover, being able to perform multi-site studies has become critical, as large consortia of meta- and mega-analyses are being formed in order to achieve a sufficiently large sample size, in view of the heterogeneity and complexity of brain aging. Toward this goal, we utilized expert-based manual hippocampus segmentations provided by the EADC-ADNI Harmonized Protocol project for manual hippocampal segmentation (HarP) (Frisoni et al., 2015), and same-day scan/re-scan images from 1.5T and 3T MRI scans in ADNI.

2. Materials and methods

2.1. MR imaging data

We used two publicly available datasets for quantitative evaluations (Table 1). The first dataset (Dataset1) was provided by the EADC-ADNI Harmonized Protocol project for manual hippocampal segmentation (HarP) (Frisoni et al., 2015). The main aim of the HarP project was to harmonize existing protocols for manual segmentation of hippocampus in order to derive standardized ground truth labels that will be used as benchmark. The HarP dataset included T1-weighted scans of 135 ADNI subjects (Mean Age: 75.014, Slice Thickness: 1.2 mm,with 44 Controls, 45 MCI, 45 AD) and left and right hippocampal labels for these scans delineated by expert raters.[4] Due to mismatch related to image format conversion for one subject scan, our final sample included a total of 134 subjects.

Our Second dataset (Dataset2) consisted of T1-weighted scans of 113 ADNI-1 subjects who underwent 1.5T and 3T T1-weighted MR Imaging at the same day or within similar dates (12 scans within 1 month; 4 scans within 2 months; 1 scan within 3 months). All subjects belonged to the patient group with mean age of 75.7. High resolution structural scans were acquired using 1.5T (TR = 2400 ms,Flip Angle = 8°, with acquisition matrix of 256 × 256 × 166, yielding voxel size of 0.9 × 0.9 × 1.2 mm) and 3T (TR = 2300 ms, Flip Angle = 8°, with acquisition matrix of 256 × 256 × 170, yielding voxel size of 1.0 × 1.0 × 1.2) MR scanners from different vendors (GE, Philips and Siemens).

2.2. ROI segmentation methods

2.2.1. Freesurfer volumetric segmentation

FreeSurfer is a software package to analyze and visualize structural neuroimaging data (Fischl et al., 2002, 2004). A widely used functionality of Freesurfer is to perform cortical and subcortical segmentation. Freesurfer segmentation consists of surface and volume-based streams with multiple processing steps. After transformation to Tailarach space, the target image is corrected for intensity inhomogeneities (bias field) and non-brain tissues are removed automatically. A high dimensional nonlinear volumetric alignment to the atlas space is performed for transferring atlas label information to the target image. The alignment includes surface deformation to optimally place gray matter (GM) / white matter (WM) and GM / cerebro-spinal fluid (CSF) boundaries, and surface inflation and registration to spherical atlas to parcellate cerebral cortex into units based on gyral and sulcal structure. The final segmentation is based on both a subject-independent probabilistic atlas, which

[4] http://www.hippocampal-protocol.net/SOPs/index.php.

was built from a training set with manual labels, and subject-specific measured values. The final label assignment at each image voxel is achieved by finding the segmentation that maximizes the probability of input signal given the prior probabilities from the training set. Freesurfer segments the brain into 34 cortical ROIs per hemisphere, which were defined based on a parcellation scheme on an inflated representation of cortex (Desikan et al., 2006).

We applied Freesurfer v5.3 and v6.0 (released in April 2017) segmentation using the default parameters (*recon-all -i T1image -s sub -sd SUBJECTS_DIR -all*) for 1.5T scans. For 3T scans, we run Freesurfer with the "−3T"' flag (*recon-all -i T1image -s sub -sd SUBJECTS_DIR −3T -all*). Freesurfer volume estimates were calculated using Desikan Killiany Atlas for all subjects using the command *asegstats2table*.

2.2.2. MUSE segmentation

The MUSE algorithm follows the multi-atlas image registration and label fusion framework. In this framework, multiple atlases with manually or semi-automatically drawn reference labels are independently registered to the target scan using deformable registration. Candidate labels from multiple registrations are fused together to calculate a consensus segmentation. Image preprocessing in MUSE included inhomogeneity correction with N4 (Tustison et al., 2010) and multi-atlas skull-stripping (Doshi et al., 2016). After skull-stripping, 11 atlas images are warped to the target image using 2 different deformable registration algorithms DRAMMS (Ou et al., 2011) and ANTS (Avants et al., 2014), and using two different regularization parameter values for these algorithms, resulting in a high variation within the ensemble, which is desirable to be able to better capture the inter-individual variability in target anatomy. The label fusion includes a local similarity term to select locally optimal atlases and an intensity term to refine the segmentation consistently with the target image's intensity profile. MUSE reference atlases included 153 anatomical ROIs. We run MUSE using the default parameters.

2.3. Quality control of segmentations

Freesurfer and MUSE segmentations were subjected to a quality control (QC) procedure that aimed to detect and exclude cases with gross errors in the segmentations, which typically happen due to major failures of atlas to target image registrations. We should note that this is different from extensive QC on individual ROIs performed in studies with relatively modest sample sizes. In this analysis, we mainly focused on establishing an automated and reproducible processing procedure, in view of large scale datasets. Accordingly, cases were flagged for exclusion only if overall segmented ROI did not overlap with the actual brain boundaries. The QC procedure involved automatic ranking of scans based on a quality score derived from ROI volumes, followed by visual inspections of the segmentations guided by the ranking score. A quality score is automatically calculated for each segmentation mask by comparing ROI volumes extracted from this mask against the distributions obtained from the entire sample. Specifically, we used PCA to reduce data dimensionality by projecting ROI values to a lower dimensional space that optimally explains the variance of the data, and we calculated the Mahalanobis distance of each sample to the sample mean in the PCA space. We used an in-house visualization tool for visual inspections, which displayed boundaries of segmented ROIs on a subset of image slices in axial, sagittal and coronal views. A binary QC flag, i.e. accept or exclude, was assigned to each segmentation, based on the consensus of two independent raters.

2.4. Determination of common ROI labels

As the reference label definitions are different for Freesurfer and MUSE, a direct comparison of ROIs is not possible. Particularly, the cortical ROIs are substantially different both in terms of naming of anatomical regions and in the details of the delineation of their boundaries. Therefore, a direct comparison of the exact volumes produced by these two methods could not be applied, hence we focused on analyzing age trends and inter-scanner consistency. In order to facilitate this process, we applied a semi-automated process to identify approximate correspondences between the two atlas ROI definitions. We computed MUSE and Freesurfer segmentations of a single reference scan and calculated the percent overlap between each pair of ROIs from the two methods. MUSE and Freesurfer ROIs were matched to each other using a greedy matching algorithm guided by the maximum percent overlap between ROI pairs. The matching algorithm allowed grouping of multiple ROIs in one set to find the optimal coverage of a single ROI in the other set. Matched ROIs have been inspected by two manual raters (DS and MH) based on the ROI names and spatial correspondences between them. Revisions have been done after mutual agreement of the two raters. The final set of ROIs included 7 deep structure regions and 31 cortical regions. A visualization of the matched ROIs between MUSE and Freesurfer is presented in Fig. 1. A complete list of the matching between MUSE and Freesurfer ROIs is given in suppl. Fig. 1.

We also computed a set of composite ROIs in order to perform comparative analyses at a coarser level. Composite ROIs included 6 lobe level regions, the "cortex" ROI that is the combination of all cortical GM ROIs, and the "sub-cortex" ROI that combines all sub-cortical GM ROIs.

2.5. Statistical analyses

In Dataset1, for which manual hippocampus labels were available, the agreement between Freesurfer and MUSE against manual labels was measured by comparing the volumes of segmented hippocampus ROIs. Note that as the delineations of the hippocampus region were different in the reference atlases for each method and in manual segmentations, a direct calculation of the overlap between segmentations, e.g. by calculating the Dice score between them, was not suitable for comparisons.

We used Pearson pairwise correlation (r) and Lin's Concordance Correlation Coefficient (CCC) to measure the reproducibility between manually segmented hippocampal labels and labels automatically segmented using MUSE and Freesurfer. Pearson correlation is a measure of linear associations of ROI volumes obtained using different segmentation methods. Concordance correlation coefficient measures the agreement between ROI volumes from two segmentation methods and penalizes differential mean shifts between them. The concordance correlation coefficient is defined as:

$$r_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where μ_x and μ_y are the means and σ_x and σ_y are the variances of the two variables, and ρ is the correlation coefficient between them.

We computed Bland-Altman plots between gold standard (Manual) ROI volumes and the ROI volumes calculated using the three automatic segmentations. A Bland-Altman plot is a graphical method that is extensively used for comparing the agreements of two measurements of the same variable and for detecting the presence of systematic bias and amount of variation between the two.

We tested the significance of age bias of volume differences between automated and ground-truth segmentations using a linear regression model for each method separately. We fitted a linear model (estimated using OLS) in R4.0.0 to predict the delta (difference in volume) between ground truth segmentations and those obtained by each of the three methods, Freesurfer v6.0, Freesurfer v5.3 and MUSE, with Age as the predictor variable. The regression model was:

$$\Delta V_k(i) = C_0 + \beta_1 * \mathrm{Age}_i$$

where k is the segmentation method for which the volume difference from manual segmentation is calculated and Age_i represent ith subject's age. P-values were corrected using FDR correction for multiple comparisons of same subject across different methods. In order to assess

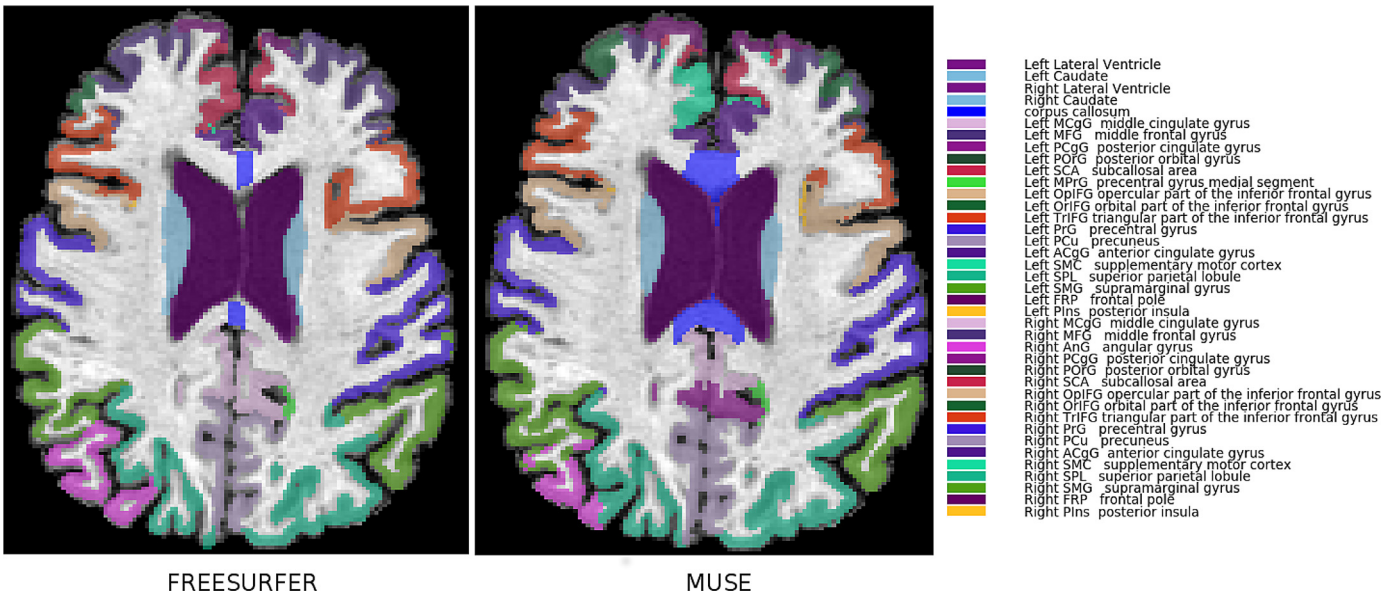

Fig. 1. ROI atlas denoting common GM ROIs in MUSE and Freesurfer.

Table 2
Left, Right & Total hippocampus volumes computed using manual ground-truth, Freesurfer and MUSE segmentations.

Method	Hippo. R	Hippo. L	Hippo. Total
Manual	2769.165 ± 590.64	2664.277 ± 593.87	5433.442 ± 1154.88
Freesurfer-v5.3	3158.656 ± 729.29	3071.671 ± 704.73	6230.328 ± 1378.62
Freesurfer-v6.0	3288.765 ± 637.11	3177.69 ± 595.06	6466.45 ± 1197.69
MUSE	3362.186 ± 606.61	3085.091 ± 585.19	6447.28 ± 1162.57
Manual – Freesurfer-v5.3(% vol diff)	14.06	15.29	14.66
Manual – Freesurfer-v6.0(% vol diff)	18.76	19.27	19.01
Manual – MUSE(% vol diff)	21.41	15.79	18.65

the significance of the correlation between volume difference and age for different diagnosis groups, we stratified subjects based on diagnosis (CN, MCI and AD) and ran separate linear regression models for each method and each diagnosis group (9 different models in total).

We used the Dataset2 for evaluating the robustness of each segmentation method to scanner variations between 1.5T and 3T scanners. For each method, we calculated Pearson pair-wise correlations (r) and concordance correlation coefficients between individual ROI volumes calculated from each pair of matching 1.5T and 3T scans. This analysis was performed for cortical and subcortical structures individually, as well as for each composite ROI.

To determine the significance of the differences between methods we used the Wilcoxon signed rank test between the ROI correlation values across scanners obtained by the two methods. We applied two independent tests, one by grouping together correlation values of all cortical ROIs, and the second by grouping correlation values of all subcortical ROIs. A non-parametric test was used to avoid any normality assumptions and the p-values were corrected for multiple comparisons.

3. Experimental results

All 132 subjects in Dataset1 were segmented using fully automated Freesurfer and MUSE pipelines. Average run time per scan for MUSE was approximately 50.4 min for individual registrations (running them in parallel) and 61.17 min for subsequent label fusion. The actual computation time depends on the number of atlases chosen for MUSE. Run time per scan for Freesurfer was approximately 9.22 h for the entire pipeline using a single thread.

In our visual inspection of segmentation results from Dataset1 and Dataset2, all MUSE segmentations passed the visual QC. For the Freesurfer segmentations, 2 cases from Dataset1 were flagged for exclusion due to gross errors (suppl. Fig. 2). The final sample included 132 subjects for Dataset1 and 113 subjects for Dataset2 after exclusion of the flagged cases.

3.1. Analysis of hippocampal volume differences

Mean volumes of right and left hippocampi calculated from automatic and manual segmentations are given in Table 2.

Fig. 2 shows the age trends of hippocampus volumes for manual and automatic segmentations. We found that MUSE and manual segmentations had a similar slope ($s_{\_MUSE}=-34.04$, $s_{\_Manual}=-31.29$), while Freesurfer segmentation had a significantly higher slope with age ($s_{\_Freesurfer\text{-}v5.3}=-52.57$,$s_{\_Freesurfer\text{-}v6.0}=-44.44$). Note that differences in intercepts between methods are expected and they are due to differences in ROI definitions. Age trends of hippocampal volumes for CN, MCI and AD subjects are shown in Fig. 3.

We verified possible bias in volume estimations using Bland Altman plots. Fig. 4 shows trends of volume differences of Freesurfer and MUSE segmentations from manual segmentations, plotted against mean hippocampus volumes.

The estimated regression line indicates a negative trend for Freesurfer, suggesting that larger hippocampus volumes lead to a larger difference in volume estimation in comparison to manual segmentations. For MUSE, regression line shows a constant difference trend for increasing hippocampus volumes. Bland Altman plots of volume differences against age are shown in Fig. 5. Similarly, both Freesurfer versions v6.0 and v5.3 has a negative slope, indicating a bias towards undersegmentation of hippocampus for older people, and hence introduction of spurious age trends. The age bias in Freesurfer segmentations was also present when the subjects were grouped by disease category (CN, MCI, AD) (suppl. Fig. 3).

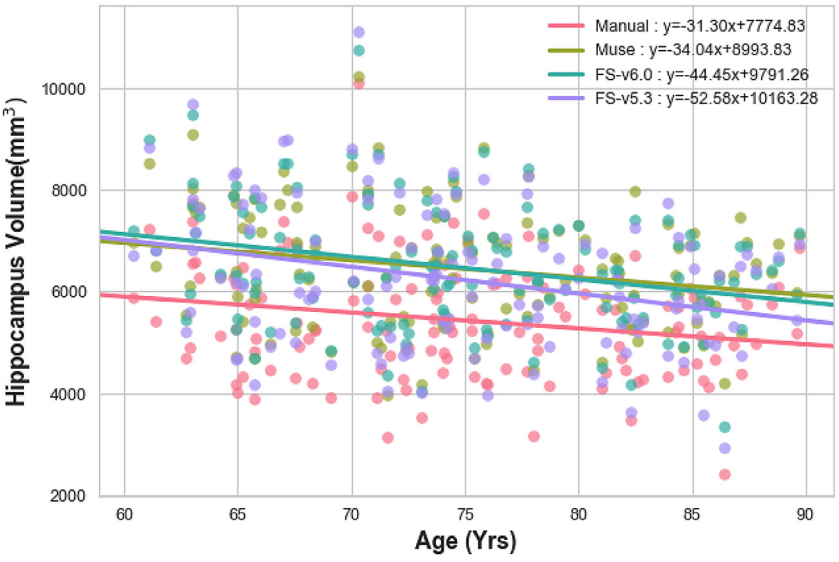

Fig. 2. Hippocampal volumes calculated from MUSE (Green), Freesurfer-v6.0(Blue), Freesurfer-v5.3(Purple) and manual(Pink) segmentations, plotted against age at scan time.

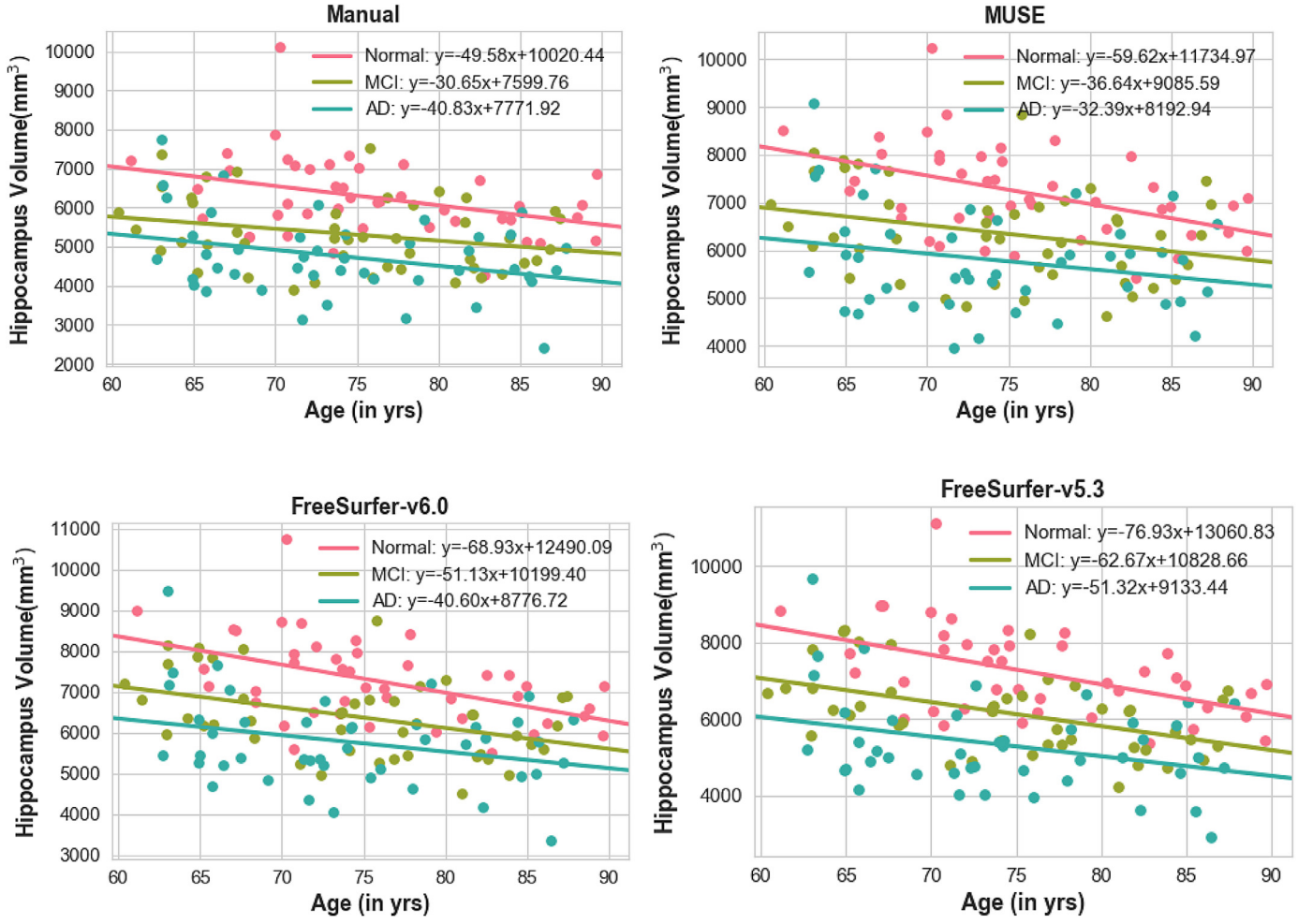

Fig. 3. Hippocampal volumes calculated from MUSE, and manual segmentation, plotted against age at scan time and grouped by disease category. Normal, MCI and AD subjects are shown with Pink, Green and Blue colors, respectively.

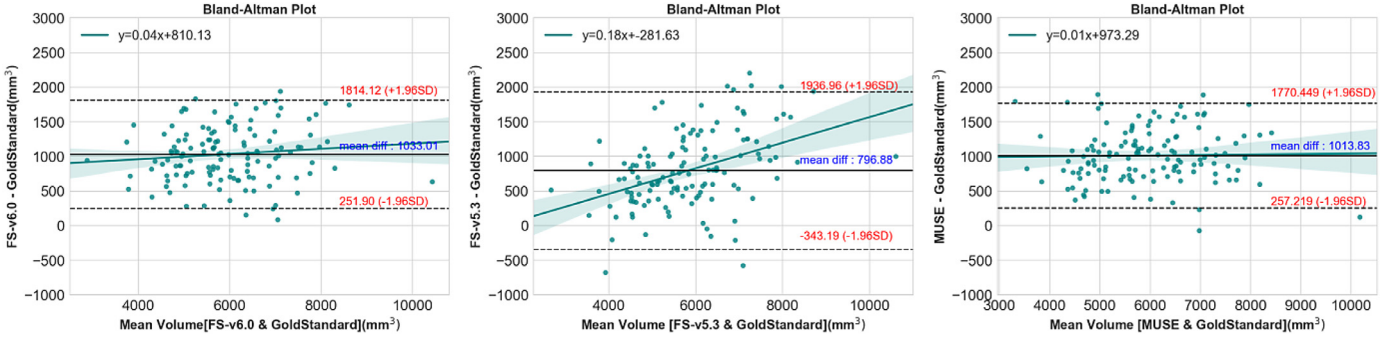

Fig. 4. Bland-Altman plots showing comparisons between hippocampal volumes calculated from Freesurfer-v6.0 and manual segmentations (left), Freesurfer-v5.3 and manual segmentations(middle) and MUSE and manual segmentations (right), against hippocampus volume. The black solid line represents the mean difference and dotted lines show upper and lower limits defined as 1.96 standard deviation from mean. X-axis and Y-axis denotes the mean values ((M1+M2)/2) and difference (M1-M2) respectively. Regression lines have been fit to the data points for each automatic method to visualize potential bias.

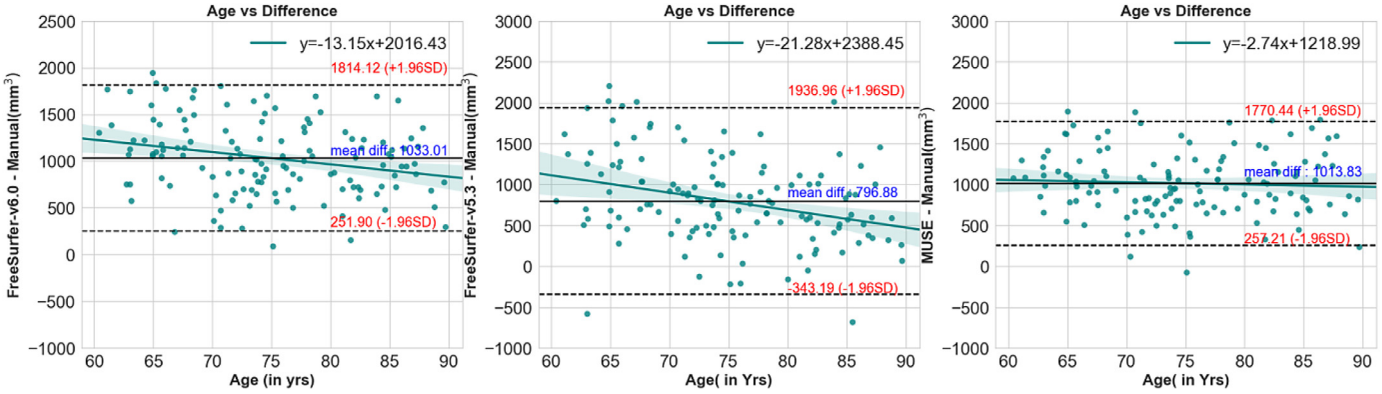

Fig. 5. Scatter plot of differences showing hippocampal volumes calculated from Freesurfer-v6.0 and manual segmentations (left), Freesurfer-v5.3 and manual segmentations (middle), and MUSE and manual segmentations (right), against subject's age at scan time.

Table 3
Age correlations of volume differences between automated and ground-truth segmentations.

	Δ_{MUSE}			$\Delta_{FS\_v5.3}$			$\Delta_{FS\_v6.0}$		
Predictors	*Estimates*	*p*	*pcorr*	*Estimates*	*p*	*pcorr*	Estimates	p	pcorr
(Intercept)	1218.99	<0.001***		2388.45	<0.001***		2016.43	<0.001***	
Age	−2.74	0.53265	0.53265	−21.28	0.00105**	0.00315**	−13.15	0.00324**	0.00487**
Observations	132			132			132		

Linear regression models showed that the observed age bias was significant in both Freesurfer-v5.3 and Freesurfer-v6.0 (Table 3). The volume difference of Freesurfer-v5.3 from the ground-truth segmentation had a higher negative slope with higher age, with a difference of 21.28 mm^3 per year of Age, 95% CI, $p < 0.05$, FDR corrected.

Results of regression models assessing the age bias for different segmentations across each diagnosis group are given in supplementary Table S1. We found that the age bias was predominantly driven by for the MCI subjects, both for Freesurfer-v5.3 and v6.0.

3.2. Reproducibility analysis across field strengths

We compared volumes of matched ROIs for 1.5T and 3T same day scan pairs segmented using the automated methods. This comparison included all individual and composite ROIs. We calculated the Pearson and Concordance correlations between 1.5T and 3T volumes for each ROI. Concordance correlation values for all individual ROIs are shown in Fig. 6.

Freesurfer-v6.0 obtained higher correlations for all ROIs compared to Freesurfer-v5.3, suggesting that the major version update in Freesurfer resulted in considerable differences in the final segmentation, improving the overall accuracy. MUSE obtained consistently higher correlation compared to Freesurfer-v5.3, while in cortical ROIs Freesurfer-v6.0 and MUSE showed comparable performance. Importantly, the differences between Freesurfer and MUSE were higher in the segmentation of the deep structures, MUSE obtaining higher correlations in all deep structures except hippocampus. Scatter plots for the volumes of selected deep structures for each method against ground truth segmentation volumes are given in Fig. 7. The distribution of correlation coefficients from all ROIs for each method is presented in Fig. 8.

The results for composite ROIs were similar, with higher correlations for MUSE in subcortical regions, while in cortical regions MUSE and Freesurfer-v6.0 had comparable reproducibility at lobe level ROIs (suppl. Fig. 4). The Wilcoxon-tests comparing Freesurfer and MUSE segmentations indicated that correlations of ROI volumes across scanners were significantly different between MUSE and both Freesurfer versions for sub-cortical ROIs. For the deep structures, the differences were significant between MUSE and Freesurfer-v5.3, but not for Freesurfer-v6.0 (Table 4).

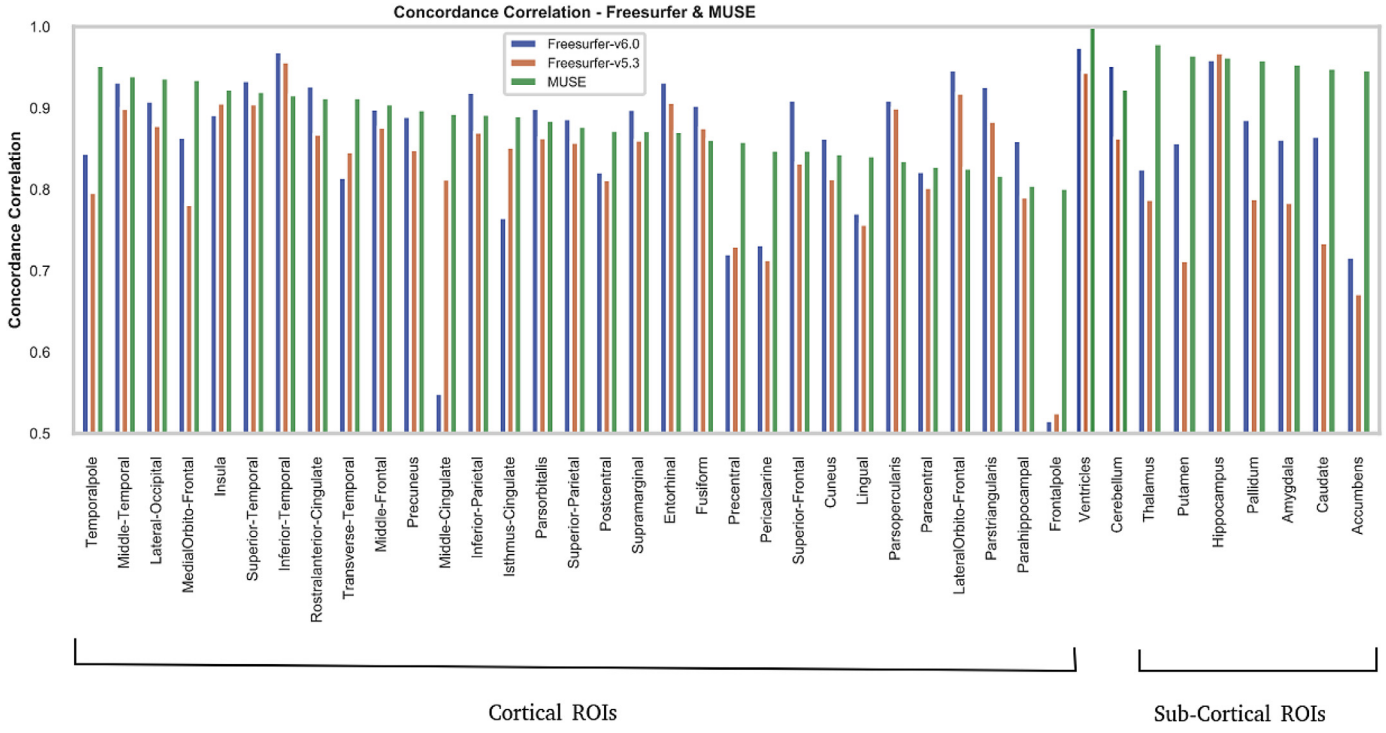

Fig. 6. Concordance correlation between ROI volumes of 1.5T and 3T ADNI scan pairs obtained using Freesurfer-v6.0, Freesurfer-v5.3 and MUSE segmentations. The ROIs are grouped as cortical and subcortical, and sorted by MUSE correlation values in decreasing order, separately for each group, to better highlight the differences.

Table 4
Mean correlation, concordance correlation and Wilcox *p*-values for segmentation of matching 1.5T and 3T ADNI scan pairs by each Freesurfer version and MUSE.

	Cortical			Sub-cortical		
	Freesurfer-v5.3	*Freesurfer-v6.0*	*MUSE*	*Freesurfer-v5.3*	*Freesurfer-v6.0*	*MUSE*
Mean Correlation	0.8867	0.8982	0.9045	0.83992	0.88656	0.97425
Mean Concordance Correlation	0.83678	0.85236	0.87805	0.78113	0.85295	0.96042
Wilcox signed rank *p*-val	0.0029	0.5052		0.03125	0.01796	
Wilcox signed rank (FDR corr)	0.0058	0.5167		0.03125	0.03125	

4. Discussion

Freesurfer is one of the most widely used tools in neuroimaging research for segmenting brain anatomy. Although Freesurfer is extensively validated and it has been used in a large number of studies, various studies have also reported high rates of failures resulting in exclusion of large number of scans, inconsistencies in segmentations and age related bias (Wenger et al., 2014; Cherbuin et al., 2009).

In recent years, methods that use the multi atlas label fusion framework have obtained state-of-the-art accuracy, showing that consensus segmentation using multiple atlases may significantly improve the segmentation accuracy, while making it also more robust to sporadic registration errors/imperfections (Iglesias and Sabuncu, 2015; Warfield, 2017).

Herein we evaluated MUSE by comparing it to Freesurfer, which is the current standard for brain anatomy segmentation. While the accuracy, reliability and reproducibility of Freesurfer has been well tested across various studies and segmentation tools, it has not been compared with current multi-atlas methods.

Importantly, Freesurfer had a major revision update in 2017 (v6.0). Over 1600 neuroimaging publications used Freesurfer for quantification of brain volumes (as listed in pubmed.gov). Most of these publications used previous versions dating before the Freesurfer-v6.0 release. Considering that previous stable version of Freesurfer (v5.3) has been used by a large number of studies in the past, we performed comparisons using both versions of Freesurfer. Our hypothesis was that MUSE would overcome some of Freesurfer's limitations. We specifically focused on two tasks, which are important in multi-site aging studies: bias of segmentation with age, and reproducibility across different field strengths. The latter was a test of inter-scanner reproducibility, an issue that is of rapidly rising significance with the emergence of large-scale meta/mega-analyses that pool data from multiple studies (Thompson et al., 2014; Davatzikos, 2018).

Automatic segmentation methods typically use a reference atlas with manually defined ROI labels and apply image registration for transferring these labels to target image space. Freesurfer is based on the registration of a single probabilistic atlas, and label assignment based on both aligned atlas probabilities and target image intensities. In contrast, multi-atlas techniques take advantage of the consensus labeling of multiple atlases. The advantage of the multi atlas approach is twofold: a) multiple atlases allow capturing a broader anatomical variation, e.g. by including atlases from subjects with different sex and age, thereby allowing the label fusion algorithm to select subject-appropriate atlases on a regional basis; and b) even when the image registration fails for one or more atlas images, the voting between multiple atlases presumably helps obtain a correct segmentation, unless there is a systematic registration error that affects a majority of the atlases. Specifically, in MUSE, two different registration algorithms were used to increase the

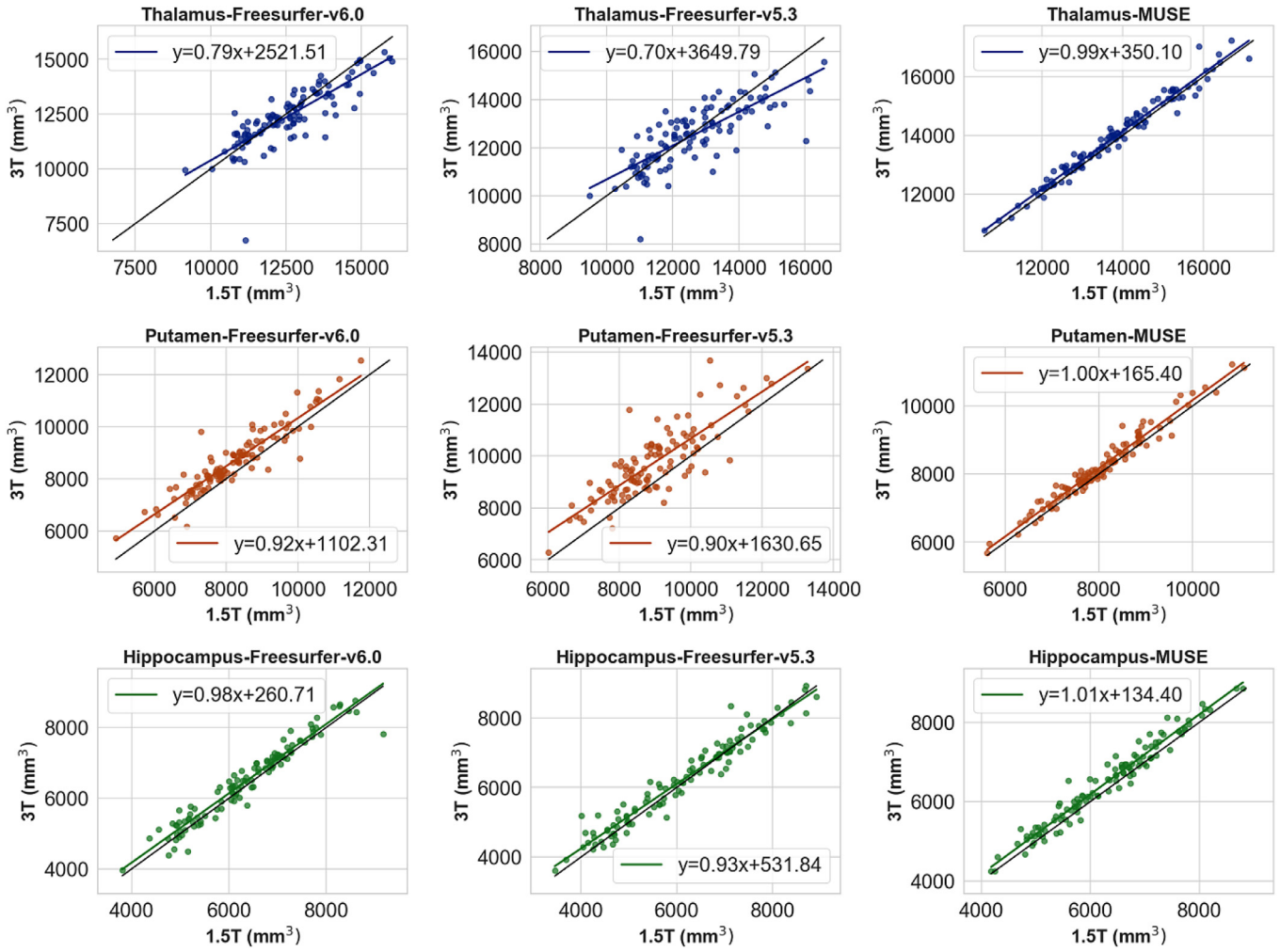

Fig. 7. Scatter plots of ROI volumes for segmentations of 1.5T and 3T ADNI scan pairs. The plots show the volumes of 3 deep structures, thalamus, putamen and hippocampus, calculated using Freesurfer-v6.0(left), Freesurfer-v5.3(middle) and MUSE (right).

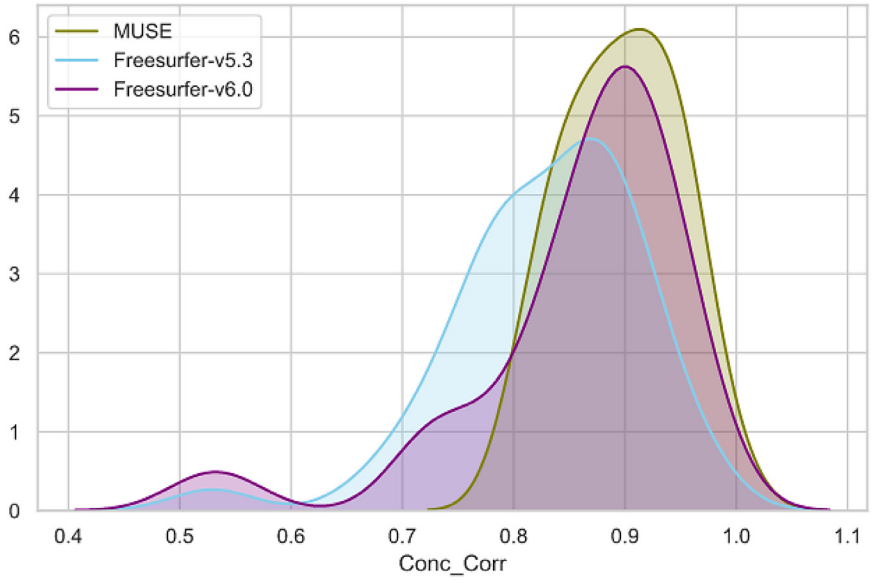

Fig. 8. Histogram showing the distribution of Concordance Correlation between 1.5T and 3T ADNI scan pairs calculated using Freesurfer-v6.0(blue), Freesurfer-v5.3 (purple) and MUSE (green).

variations within the ensemble, and a local similarity ranking strategy was used to give more weights to atlases locally more similar to the target scan in the fusion.

We found that hippocampus segmentations by both Freesurfer-v6.0 and Freesurfer-v5.3 showed a bias towards over-segmentation of larger hippocampi, and under-segmentation of hippocampi in smaller or older individuals, although this age bias was reduced with the newer version. On the other hand, MUSE segmentations were more consistent with ground-truth segmentations, without bias with subject age or hippocampus volume. This is an important finding, when evaluating age effects on brain volumes. Our results suggest that Freesurfer introduces spurious age effects, which can obviously lead to false biological interpretations. One reason could be that Freesurfer does not use the population-based specific template.

In our reproducibility analyses, we found that MUSE obtained consistently higher correlations between ROI volumes calculated across the two scanners in comparison to Freesurfer 5.3. Importantly, we also found that Freesurfer 6.0 segmentations showed consistent improvement in all ROIs compared to Freesurfer-v5.3. While Freesurfer 6.0 was comparable to MUSE for cortical ROIs, in deep structures MUSE obtained very consistent segmentations across subjects with higher correlations compared to both Freesurfer versions. This finding is consistent with previously reported state-of-the-art accuracy of MAS methods in segmentation of deep structures, and could be explained by the advantage of using a model based (i.e. a-priori defined ground-truth ROI labels) consensus labeling in these regions where the tissue contrast alone is not informative enough to guide the segmentation.

High number of Freesurfer exclusions based on detailed visual verification of individual ROIs was previously reported (Mccarthy 2015; Zandifar et al., 2017). Such an extensive QC was out of scope of our analysis that focused on mostly automated processing in view of largescale datasets. In our QC for detecting gross errors, MUSE outperformed Freesurfer in terms of overall failure rates. While ~1% of all segmentations (2 scans from 247 in total) were excluded due to Freesurfer failures based on our case by case visual QC of general segmentation quality, all MUSE segmentations have obtained a positive QC result. The robustness of MUSE was expected, as multiple atlases provide various representations of the anatomy, while the label fusion of multiple warped atlases allows the method to correct the effect of individual registrations that failed, unless the failure is not systematic to a majority of the atlases.

This work has also some limitations. A major problem for a systematic comparison is the limited availability of ground truth segmentations. Manual segmentation of anatomical ROIs is a difficult and time consuming task. For this reason, there are very few datasets that provide manually segmented ROI labels. Also, because Freesurfer and MUSE use their own atlas sets with different ROI label denotations, a direct comparison of the two methods is not possible. MUSE uses a set of 35 scans and their semi-automatically segmented ROI labels as reference atlases. We did not prefer to use these scans in our comparisons, as this would be biased towards MUSE, even with cross-validation. Also, our experiments were limited to comparisons between Freesurfer and the multi atlas segmentation approach. In recent years deep learning methods obtained state of the art accuracy in various problems in neuroimaging. While a comparison to more recent deep learning methods for segmentation would be very informative, this is out of the scope of this paper.

Our comparative evaluations have shown that MUSE, a multi-atlas ROI segmentation method, can help alleviate some of the limitations of Freesurfer and related methods, by virtue of leveraging multiple atlases, registration methods and parameters, thereby offering both the advantages of a consensus-based methods and of regional adaptivity of the atlases to the target anatomy. Critically, MUSE also displayed significantly higher inter-scanner consistency, thereby offering promise that multi-site, multi-study meta/mega-analyses can be performed more accurately. Given these favorable results and increasing availability of parallel and cloud computing capacities, multi-atlas segmentation has a great potential of becoming the standard approach for segmentation of brain images in population studies and in clinical applications.

Acknowledgments

This work was supported by the National Institute on Aging (grant number 1RF1AG054409), the National Institute of Mental Health (grant number 5R01MH112070), National Institutes of Health (grant number 1RF1AG059869) and National Institutes of Health (grant number 75N95019C00022).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117248.

References

Apostolova, L.G., Dutton, R.A., Dinov, I.D., Hayashi, K.M., Toga, A.W., Cummings, J.L., Thompson, P.M., 2006. Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. Arch. Neurol. 63, 693–699. doi:10.1001/archneur.63.5.693.

Asman, A., Alireza Akhondi-Asl, Wang, H., Tustison, N., Avants, B., Warfield, S.K., Landman, B., 2013. Miccai 2013 segmentation algorithms, theory and applications (SATA) challenge results summary. Presented at the MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications (SATA).

Avants, B.B., Tustison, N.J., Stauffer, M., Song, G., Wu, B., Gee, J.C., 2014. The Insight ToolKit image registration framework. Front Neuroinform 8. doi:10.3389/fninf.2014.00044.

Bakkour, A., Morris, J.C., Wolk, D.A., Dickerson, B.C., 2013. The effects of aging and Alzheimer's disease on cerebral cortical anatomy: specificity and differential relationships with cognition. Neuroimage 76, 332–344. doi:10.1016/j.neuroimage.2013.02.059.

Bonilha, L., Molnar, C., Horner, M.D., Anderson, B., Forster, L., George, M.S., Nahas, Z., 2008. Neurocognitive deficits and prefrontal cortical atrophy in patients with schizophrenia. Schizophr Res 101, 142–151. doi:10.1016/j.schres.2007.11.023.

Brewer, J.B., Magda, S., Airriess, C., Smith, M.E., 2009. Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in Alzheimer disease. AJNR Am J Neuroradiol 30, 578–580. doi:10.3174/ajnr.A1402.

Charil, A., Dagher, A., Lerch, J.P., Zijdenbos, A.P., Worsley, K.J., Evans, A.C., 2007. Focal cortical atrophy in multiple sclerosis: relation to lesion load and disability. Neuroimage 34, 509–517. doi:10.1016/j.neuroimage.2006.10.006.

Cherbuin, N., Anstey, K.J., Réglade-Meslin, C., Sachdev, P.S., 2009. In vivo hippocampal measurement and memory: a comparison of manual tracing and automated segmentation in a large community-based sample. PLoS ONE 4, e5265. doi:10.1371/journal.pone.0005265.

Davatzikos, C., 2018. BRAIN AGING HETEROGENEITY ELUCIDATED VIA MACHINE LEARNING: THE MULTI-SITE ISTAGING DIMENSIONAL NEUROIMAGING REFERENCE SYSTEM. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 14, P1476–P1477. doi:10.1016/j.jalz.2018.06.2505.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980. doi:10.1016/j.neuroimage.2006.01.021.

Dicks, E., Vermunt, L., van der Flier, W.M., Visser, P.J., Barkhof, F., Scheltens, P., Tijms, B.M.Alzheimer's Disease Neuroimaging Initiative, 2019. Modeling grey matter atrophy as a function of time, aging or cognitive decline show different anatomical patterns in Alzheimer's disease. Neuroimage Clin 22, 101786. doi:10.1016/j.nicl.2019.101786.

Doshi, J., Erus, G., Ou, Y., Resnick, S.M., Gur, R.C., Gur, R.E., Satterthwaite, T.D., Furth, S., Davatzikos, C.Alzheimer's Neuroimaging Initiative, 2016. MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. Neuroimage 127, 186–195. doi:10.1016/j.neuroimage.2015.11.073.

Ferreira, D., Hansson, O., Barroso, J., Molina, Y., Machado, A., Hernández-Cabrera, J.A., Muehlboeck, J.-S., Stomrud, E., Nägga, K., Lindberg, O., Ames, D., Kalpouzos, G., Fratiglioni, L., Bäckman, L., Graff, C., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Ahlström, H., Lind, L., Larsson, E.-M., Wahlund, L.-O., Simmons, A., Westman, E.the AddNeuroMed consortium, for the Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) research group, 2017. The interactive effect of demographic and clinical factors on hippocampal volume: A multicohort study on 1958 cognitively normal individuals. Hippocampus 27, 653–667. doi:10.1002/hipo.22721.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355. doi:10.1016/s0896-6273(02)00569-x.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B.,

Dale, A.M., 2004. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14, 11–22. doi:10.1093/cercor/bhg087.

Frisoni, G.B., Jack, C.R., Bocchetta, M., Bauer, C., Frederiksen, K.S., Liu, Y., Preboske, G., Swihart, T., Blair, M., Cavedo, E., Grothe, M.J., Lanfredi, M., Martinez, O., Nishikawa, M., Portegies, M., Stoub, T., Ward, C., Apostolova, L.G., Ganzola, R., Wolf, D., Barkhof, F., Bartzokis, G., DeCarli, C., Csernansky, J.G., deToledo-Morrell, L., Geerlings, M.I., Kaye, J., Killiany, R.J., Lehéricy, S., Matsuda, H., O'Brien, J., Silbert, L.C., Scheltens, P., Soininen, H., Teipel, S., Waldemar, G., Fellgiebel, A., Barnes, J., Firbank, M., Gerritsen, L., Henneman, W., Malykhin, N., Pruessner, J.C., Wang, L., Watson, C., Wolf, H., deLeon, M., Pantel, J., Ferrari, C., Bosco, P., Pasqualetti, P., Duchesne, S., Duvernoy, H., Boccardi, M.EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Volumetry and for the Alzheimer's Disease Neuroimaging Initiative, 2015. The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimers Dement 11, 111–125. doi:10.1016/j.jalz.2014.05.1756.

Giorgio, A., De Stefano, N., 2013. Clinical use of brain volumetry. J Magn Reson Imaging 37, 1–14. doi:10.1002/jmri.23671.

Goldstein, J.M., Goodman, J.M., Seidman, L.J., Kennedy, D.N., Makris, N., Lee, H., Tourville, J., Caviness, V.S., Faraone, S.V., Tsuang, M.T., 1999. Cortical abnormalities in schizophrenia identified by structural magnetic resonance imaging. Arch. Gen. Psychiatry 56, 537–547. doi:10.1001/archpsyc.56.6.537.

Habes, M., Toledo, J.B., Resnick, S.M., Doshi, J., Van der Auwera, S., Erus, G., Janowitz, D., Hegenscheid, K., Homuth, G., Völzke, H., Hoffmann, W., Grabe, H.J., Davatzikos, C., 2016. Relationship between APOE Genotype and Structural MRI Measures throughout Adulthood in the Study of Health in Pomerania Population-Based Cohort. AJNR Am J Neuroradiol 37, 1636–1642. doi:10.3174/ajnr.A4805.

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: A survey. Med Image Anal 24, 205–219. doi:10.1016/j.media.2015.06.012.

Janowitz, D., Schwahn, C., Borchardt, U., Wittfeld, K., Schulz, A., Barnow, S., Biffar, R., Hoffmann, W., Habes, M., Homuth, G., Nauck, M., Hegenscheid, K., Lotze, M., Völzke, H., Freyberger, H.J., Debette, S., Grabe, H.J., 2014. Genetic, psychosocial and clinical factors associated with hippocampal volume in the general population. Transl Psychiatry 4, e465. doi:10.1038/tp.2014.102.

Keller, S.S., Gerdes, J.S., Mohammadi, S., Kellinghaus, C., Kugel, H., Deppe, K., Ringelstein, E.B., Evers, S., Schwindt, W., Deppe, M., 2012. Volume estimation of the thalamus using freesurfer and stereology: consistency between methods. Neuroinformatics 10, 341–350. doi:10.1007/s12021-012-9147-0.

Kikinis, Z., Fallon, J.H., Niznikiewicz, M., Nestor, P., Davidson, C., Bobrow, L., Pelavin, P.E., Fischl, B., Yendiki, A., McCarley, R.W., Kikinis, R., Kubicki, M., Shenton, M.E., 2010. Gray matter volume reduction in rostral middle frontal gyrus in patients with chronic schizophrenia. Schizophr. Res. 123, 153–159. doi:10.1016/j.schres.2010.07.027.

McCarthy, C.S., Ramprashad, A., Thompson, C., Botti, J.-A., Coman, I.L., Kates, W.R., 2015. A comparison of FreeSurfer-generated data with and without manual intervention. Front Neurosci 9, 379. doi:10.3389/fnins.2015.00379.

Messina, D., Cerasa, A., Condino, F., Arabia, G., Novellino, F., Nicoletti, G., Salsone, M., Morelli, M., Lanza, P.L., Quattrone, A., 2011. Patterns of brain atrophy in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. Parkinsonism Relat. Disord. 17, 172–176. doi:10.1016/j.parkreldis.2010.12.010.

Mulder, E.R., de Jong, R.A., Knol, D.L., van Schijndel, R.A., Cover, K.S., Visser, P.J., Barkhof, F., Vrenken, H.Alzheimer's Disease Neuroimaging Initiative, 2014. Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 92, 169–181. doi:10.1016/j.neuroimage.2014.01.058.

Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C., 2011. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. Med Image Anal 15, 622–639. doi:10.1016/j.media.2010.07.002.

Raz, N., Ghisletta, P., Rodrigue, K.M., Kennedy, K.M., Lindenberger, U., 2010. Trajectories of brain aging in middle-aged and older adults: regional and individual differences. Neuroimage 51, 501–511. doi:10.1016/j.neuroimage.2010.03.020.

Rohrer, J.D., Lashley, T., Schott, J.M., Warren, J.E., Mead, S., Isaacs, A.M., Beck, J., Hardy, J., de Silva, R., Warrington, E., Troakes, C., Al-Sarraj, S., King, A., Borroni, B., Clarkson, M.J., Ourselin, S., Holton, J.L., Fox, N.C., Revesz, T., Rossor, M.N., Warren, J.D., 2011. Clinical and neuroanatomical signatures of tissue pathology in frontotemporal lobar degeneration. Brain 134, 2565–2581. doi:10.1093/brain/awr198.

Sabuncu, M.R., Desikan, R.S., Sepulcre, J., Yeo, B.T.T., Liu, H., Schmansky, N.J., Reuter, M., Weiner, M.W., Buckner, R.L., Sperling, R.A., Fischl, B.Alzheimer's Disease Neuroimaging Initiative, 2011. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. Arch. Neurol. 68, 1040–1048. doi:10.1001/archneurol.2011.167.

Satterthwaite, T.D., Wolf, D.H., Calkins, M.E., Vandekar, S.N., Erus, G., Ruparel, K., Roalf, D.R., Linn, K.A., Elliott, M.A., Moore, T.M., Hakonarson, H., Shinohara, R.T., Davatzikos, C., Gur, R.C., Gur, R.E., 2016. Structural Brain Abnormalities in Youth With Psychosis Spectrum Symptoms. JAMA Psychiatry 73, 515–524. doi:10.1001/jamapsychiatry.2015.3463.

Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., Wright, M.J., Martin, N.G., Agartz, I., Alda, M., Alhusaini, S., Almasy, L., Almeida, J., Alpert, K., Andreasen, N.C., Andreassen, O.A., Apostolova, L.G., Appel, K., Armstrong, N.J., Aribisala, B., Bastin, M.E., Bauer, M., Bearden, C.E., Bergmann, Ø., Binder, E.B., Blangero, J., Bockholt, H.J., Bøen, E., Bois, C., Boomsma, D.I., Booth, T., Bowman, I.J., Bralten, J., Brouwer, R.M., Brunner, H.G., Brohawn, D.G., Buckner, R.L., Buitelaar, J., Bulayeva, K., Bustillo, J.R., Calhoun, V.D., Cannon, D.M., Cantor, R.M., Carless, M.A., Caseras, X., Cavalleri, G.L., Chakravarty, M.M., Chang, K.D., Ching, C.R.K., Christoforou, A., Cichon, S., Clark, V.P., Conrod, P., Coppola, G., Crespo-Facorro, B., Curran, J.E., Czisch, M., Deary, I.J., de Geus, E.J.C., den Braber, A., Delvecchio, G., Depondt, C., de Haan, L., de Zubicaray, G.I., Dima, D., Dimitrova, R., Djurovic, S., Dong, H., Donohoe, G., Duggirala, R., Dyer, T.D., Ehrlich, S., Ekman, C.J., Elvsåshagen, T., Emsell, L., Erk, S., Espeseth, T., Fagerness, J., Fears, S., Fedko, I., Fernández, G., Fisher, S.E., Foroud, T., Fox, P.T., Francks, C., Frangou, S., Frey, E.M., Frodl, T., Frouin, V., Garavan, H., Giddaluru, S., Glahn, D.C., Godlewska, B., Goldstein, R.Z., Gollub, R.L., Grabe, H.J., Grimm, O., Gruber, O., Guadalupe, T., Gur, R.E., Gur, R.C., Göring, H.H.H., Hagenaars, S., Hajek, T., Hall, G.B., Hall, J., Hardy, J., Hartman, C.A., Hass, J., Hatton, S.N., Haukvik, U.K., Hegenscheid, K., Heinz, A., Hickie, I.B., Ho, B.-C., Hoehn, D., Hoekstra, P.J., Hollinshead, M., Holmes, A.J., Homuth, G., Hoogman, M., Hong, L.E., Hosten, N., Hottenga, J.-J., Hulshoff Pol, H.E., Hwang, K.S., Jack, C.R., Jenkinson, M., Johnston, C., Jönsson, E.G., Kahn, R.S., Kasperaviciute, D., Kelly, S., Kim, S., Kochunov, P., Koenders, L., Krämer, B., Kwok, J.B.J., Lagopoulos, J., Laje, G., Landen, M., Landman, B.A., Lauriello, J., Lawrie, S.M., Lee, P.H., Le Hellard, S., Lemaître, H., Leonardo, C.D., Li, C., Liberg, B., Liewald, D.C., Liu, X., Lopez, L.M., Loth, E., Lourdusamy, A., Luciano, M., Macciardi, F., Machielsen, M.W.J., MacQueen, G.M., Malt, U.F., Mandl, R., Manoach, D.S., Martinot, J.-L., Matarin, M., Mather, K.A., Mattheisen, M., Mattingsdal, M., Meyer-Lindenberg, A., McDonald, C., McIntosh, A.M., McMahon, F.J., McMahon, K.L., Meisenzahl, E., Melle, I., Milaneschi, Y., Mohnke, S., Montgomery, G.W., Morris, D.W., Moses, E.K., Mueller, B.A., Muñoz Maniega, S., Mühleisen, T.W., Müller-Myhsok, B., Mwangi, B., Nauck, M., Nho, K., Nichols, T.E., Nilsson, L.-G., Nugent, A.C., Nyberg, L., Olvera, R.L., Oosterlaan, J., Ophoff, R.A., Pandolfo, M., Papalampropoulou-Tsiridou, M., Papmeyer, M., Paus, T., Pausova, Z., Pearlson, G.D., Penninx, B.W., Peterson, C.P., Pfennig, A., Phillips, M., Pike, G.B., Poline, J.-B., Potkin, S.G., Pütz, B., Ramasamy, A., Rasmussen, J., Rietschel, M., Rijpkema, M., Risacher, S.L., Roffman, J.L., Roiz-Santiañez, R., Romanczuk-Seiferth, N., Rose, E.J., Royle, N.A., Rujescu, D., Ryten, M., Sachdev, P.S., Salami, A., Satterthwaite, T.D., Savitz, J., Saykin, A.J., Scanlon, C., Schmaal, L., Schnack, H.G., Schork, A.J., Schulz, S.C., Schür, R., Seidman, L., Shen, L., Shoemaker, J.M., Simmons, A., Sisodiya, S.M., Smith, C., Smoller, J.W., Soares, J.C., Sponheim, S.R., Sprooten, E., Starr, J.M., Steen, V.M., Strakowski, S., Strike, L., Sussmann, J., Sämann, P.G., Teumer, A., Toga, A.W., Tordesillas-Gutierrez, D., Trabzuni, D., Trost, S., Turner, J., Van den Heuvel, M., van der Wee, N.J., van Eijk, K., van Erp, T.G.M., van Haren, N.E.M., van 't Ent, D., van Tol, M.-J., Valdés Hernández, M.C., Veltman, D.J., Versace, A., Völzke, H., Walker, R., Walter, H., Wang, L., Wardlaw, J.M., Weale, M.E., Weiner, M.W., Wen, W., Westlye, L.T., Whalley, H.C., Whelan, C.D., White, T., Winkler, A.M., Wittfeld, K., Woldehawariat, G., Wolf, C., Zilles, D., Zwiers, M.P., Thalamuthu, A., Schofield, P.R., Freimer, N.B., Lawrence, N.S., Drevets, W., 2014. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging and Behavior. doi:10.1007/s11682-013-9269-5.

Tian, Q., Bair, W.-N., Resnick, S.M., Bilgel, M., Wong, D.F., Studenski, S.A., 2018. β-amyloid deposition is associated with gait variability in usual aging. Gait Posture 61, 346–352. doi:10.1016/j.gaitpost.2018.02.002.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29, 1310–1320. doi:10.1109/TMI.2010.2046908.

Wee, P., Wang, Z., 2017. Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. Cancers (Basel) 9. doi:10.3390/cancers9050052.

Wenger, E., Mårtensson, J., Noack, H., Bodammer, N.C., Kühn, S., Schaefer, S., Heinze, H.-J., Düzel, E., Bäckman, L., Lindenberger, U., Lövdén, M., 2014. Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains. Hum Brain Mapp 35, 4236–4248. doi:10.1002/hbm.22473.

Wierenga, L., Langen, M., Ambrosino, S., van Dijk, S., Oranje, B., Durston, S., 2014. Typical development of basal ganglia, hippocampus, amygdala and cerebellum from age 7 to 24. Neuroimage 96, 67–72. doi:10.1016/j.neuroimage.2014.03.072.

Zandifar, A., Fonov, V., Coupé, P., Pruessner, J., Collins, D.L.Alzheimer's Disease Neuroimaging Initiative, 2017. A comparison of accurate automatic hippocampal segmentation methods. Neuroimage 155, 383–393. doi:10.1016/j.neuroimage.2017.04.018.