Psychiatry Research: Neuroimaging 280 (2018) 39–47

Contents lists available at ScienceDirect

Psychiatry Research: Neuroimaging

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

Can measuring hippocampal atrophy with a fully automatic method be substantially less noisy than manual segmentation over both 1 and 3 years?

Keith S. Cover[a,*], Ronald A. van Schijndel[a], Paolo Bosco[b], Soheil Damangir[c], Alberto Redolfi[d], Alzheimer's Disease Neuroimaging Initiative[1]

[a] Amsterdam University Medical Center, Amsterdam, The Netherlands
[b] National Institute for Nuclear Physics, Pisa, Italy
[c] Karolinska Institutet, Stockholm, Sweden
[d] IRCCS San Giovanni di Dio Fatebenefratelli, Brescia, Italy

ARTICLE INFO

Keywords:
Hippocampus
Atrophy
Manual segmentation
Automatic segmentation
Magnetic resonance imaging
Alzheimer's disease
Binomial

ABSTRACT

To quantify the "segmentation noise" of several widely used fully automatic methods for measuring longitudinal hippocampal atrophy in Alzheimer's disease and compare the results to the segmentation noise of manual segmentation over both 1 and 3 years. The segmentation noise of 5 longitudinal hippocampal atrophy measurement methods was quantified, including checking its Gaussianity, using 264 subjects from the ADNI1 back-to-back (BTB) data set over both 1 year and 3 year intervals. The segmentation methods were FreeSurfer 5.3.0 both cross sectional and longitudinal, FreeSurfer 6.0.0 longitudinal, MAPS-HBSI and FSL/FIRST 5.0.8. The BTB manual segmentation of 75 ADNI subjects from a previous study provided the manual distributions for comparison. All methods, including the manual segmentation, violated the Gaussianity assumption. Two methods, FreeSurfer 6.0.0 and MAPS-HBSI, had a segmentation noise substantially less than a surrogate for manual segmentation. FreeSurfer 5.3.0 longitudinal was confirmed as a surrogate for manual segmentation. The violation of the Gaussian assumption by the segmentation methods assessed, including manual, suggests results of previous studies that assumed Gaussian statistics without confirmation may need review. Fully automatic FreeSurfer 6.0.0 and MAPS-HBSI both have lower segmentation noise than manual requiring less than two thirds of the subjects to detect the same treatment effect.

1. Introduction

Hippocampal atrophy is the amount of shrinkage of the hippocampus from one time point to the next. It can be measured with noninvasive MRI and is a widely validated surrogate outcome for Alzheimer's disease (AD) trials (Frisoni et al., 2010). It has been shown to be one of the first observable characteristics of AD (Bobinski et al., 1996). It also accelerates before the translation to clinical dementia (Jack et al., 2011) as part of the AD pathology cascade (Jack et al., 2010). Analysis of the images from the ADNI1 study found the median annualized atrophy rates were 1.5% (healthy controls (HC)), 2.4% (mildly cognitively impaired (MCI)) and 5.1% (AD) (Cover et al., 2016).

Many software methods are available to fully automatically measure hippocampal atrophy directly (longitudinal measurement) or indirectly by measuring the change in volume between two time points (cross sectional measurement). Segmentation methods include FreeSurfer (Fischl et al., 2012), FSL/FIRST (Patenaude et al., 2011) and MAPS-HBSI (Leung et al., 2010). Several fully automatic segmentation methods also have government approval for clinical use including NeuroReader (Ahdidan et al., 2015; NeuroReader, 2016), LEAP (Woltz et al., 2014; LEAP, 2016) and NeuroQuant (Ochs et al., 2015; NeuroQuant, 2016).

Correctly assessing the performance of fully automatic segmentation methods is particularly challenging when the methods perform better than the gold standard of manual segmentation. Recently, we reported that a fully automatic segmentation method (MAPS-HBSI) (Leung et al.,

* Corresponding author.
E-mail address: keith@kscover.ca (K.S. Cover).

[1] 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.pscychresns.2018.06.011
Received 19 February 2018; Received in revised form 26 June 2018; Accepted 27 June 2018
Available online 09 August 2018
0925-4927/

2010) had substantially lower segmentation noise than manual segmentation (Cover et al., 2016).

Almost all studies assessing the segmentation noise of the hippocampus - and other structures such as the whole brain or cortical thickness - have used parametric statistics which assume Gaussian distributions - such as the mean, standard deviation and interclass correlation coefficient. Gaussian distributions are also referred to as normal distributions. There are only a few exceptions (Smith et al., 2007; Cover et al., 2011; Mulder et al., 2014; Cover et al., 2016; Opfer et al., 2016). The validity of the parametric statistics rests on the segmentation noise having a Gaussian distribution. While the segmentation noise of the whole brain has been shown to violate the Gaussian assumption (Cover et al., 2011), no study in the literature has checked whether any of the segmentation methods measuring hippocampal atrophy has a Gaussian noise distribution. Rarely has the potential impact of non-Gaussian distributions on parametric statistical calculations such as sample size been considered.

The current study focuses on assessing the segmentation noise - as measured by the back-to-back (BTB) reproducibility - of hippocampal atrophy measuring methods that have lower segmentation noise than that of the manual method. The segmentation noise for all methods is analyzed with both parametric and robust statistical methods. Also, the Gaussianity of the segmentation noise distributions is checked to determine if robust statistics are required. In addition, the segmentation noise of FreeSurfer 5.3.0 longitudinal is compared to the segmentation noise of manual measurements to confirm FreeSurfer 5.3.0 longitudinal is a suitable surrogate for the noise of manual segmentation. Finally, the segmentation noise over 1 and 3 years for all methods is compared to the surrogate for manual segmentation noise.

2. Methods

2.1. Dataset

The ADNI1 data set is widely used in studies of the reproducibility of structural measures including the hippocampus (Cover et al., 2011; Mulder et al., 2014; Ochs et al., 2015; Ahdidan et al., 2015; Chincarini et al., 2016; Cover et al., 2016). The 1.5T T1-weighted MRI scans were selected from the ADNI database and downloaded in their original unprocessed DICOM format. A total of 264 subjects are selected that had two BTB scans at baseline, 1 year and 3 years for a total of $6 \times 264 = 1{,}584$ image volumes. Supplemental table S1 has a complete listing of the subjects used including exact identification of the image volume. The 264 subjects in the current study are a subset of the 562 ADNI1 BTB 1.5T subjects in a previous study (Cover et al., 2016). Only 264 of the 562 subjects also had BTB scans at 3 years in ADNI1 therefore only 264 subjects are used in the current study.

The ADNI1 study acquired the MRI sequence twice during each patient visit. The subjects did not leave the MRI between MPRAGEs and often the second MPRAGE started within a few second of the completion of the first. While the images from only one MPRAGE sequence at each patient visit are needed to calculate the hippocampal atrophy, the second MPRAGE provides excellent data to make noise measurements. The two MPRAGE sequences are referred to as BTB, rather than scan-rescan, because they were acquired without the patient leaving the MRI scanner. The topic of the current paper is the noise of the segmentation methods. It is a reasonable assumption that all the segmentations methods in the current paper are relatively accurate as they are widely used. Thus, the accuracy of the segmentation methods is beyond the scope of the current paper.

The 264 subjects in the current study contained 120 healthy controls (HC), 143 mildly cognitively impaired (MCI) and 1 probable AD as classified by the ADNI1 study The low number of probable AD subjects is likely due to the higher probability of probable AD subjects dropping from the study over the first 3 years. Table 1 provides descriptive statistics of the 264 subjects.

Table 1
Descriptive statistics of the 264 subjects from the ADNI1 included in the current study. The interquartiles are in brackets.

Cohort	Status	Sample size	M/F	Age (Baseline)
3 year fully automatic	HC	120	66/54	75.0 (72.0, 78.5)
	MCI	143	100/43	74.1 (70.6, 80,5)
	AD	1	1/0	78.4
	Combined	264	167/97	74.4 (71.5, 79.5)
1 year manual	HC	19	11/8	76.5 (72.1, 79.6)
	MCI	38	25/13	73.7 (70.7, 77.9)
	AD	18	7/11	74.1 (69.4, 78.4)
	Combined	75	43/32	74.1 (70.7, 77.9)

As no manual segmentation was performed as part of the current study, 75 ADNI1 BTB 1.5T subjects used in prior studies (Mulder et al., 2014; Cover et al., 2016) were also included in the current study to provide some statistics on the performance of manual segmentation. While the 75 subjects are also a subset of the 562 subjects used in a previous study (Cover et al., 2016), as are the 264 subjects mentioned above, only 40 of the subjects were common to both.

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

2.2. Statistical Analysis

A detailed description of the statistical analysis for BTB atrophy measurement has been presented previously (Cover et al., 2016). Fig. 1 of the current paper provides the steps to calculate the BTB differences over 1 year and 3 years. Additional details of the calculations follow.

The amount of atrophy - as measured by the percentage volume change (PVC) - from baseline (V_A) to year 1 or year 3 (V_B) was calculated by the equation $100*(V_B\text{-}V_A)/V_A$. For each subject there were 8 PVCs - 2 for the left and right hippocampus, 2 for the 1 year and 3 years intervals and 2 for the BTB acquisition. The BTB differences were calculated by subtracting the PVCs of the first acquired image volume of a subject visit from that of the second. Consequently, there were 4 BTB differences for each subject - one each for the left and right hippocampus and one each for the 1 year and 3 year intervals. As a result, there were 4 BTB difference distributions for each method.

A variety of statistics were calculated for each BTB difference distribution. All statistics were calculated from the absolute values of the BTB differences. The statistics included the maximum, minimum, median (MDBTBD), mean (MNBTBD) and standard deviation (SDBTBD). The value of the mean subtracted off before calculating the standard deviation was assumed to be zero. The number of BTB differences in each distribution is also listed so the number of times each method failed to yield a BTB difference can be determined.

Three different statistical tests were used to test the Gaussianity of the BTB difference distributions. Two of the tests, the Anderson-Darling and the Shapiro-Wilk tests, tested general properties of Gaussianity. The third test was tailored to whether the distributions had too many outliers for a Gaussian distribution.

The tailored test is based on the ratio of SDBTBD and MDBTBD, two measures of the spread of BTB distributions used in the literature. For an ideal Gaussian distribution the ratio of SDBTBD/MDBTBD is 1.3654. However, as the standard deviation of a distribution is more sensitive to the distribution's shoulders the ratio increases as the shoulders get larger. To calculate the *p*-value for a range of ratios, 10,000,000

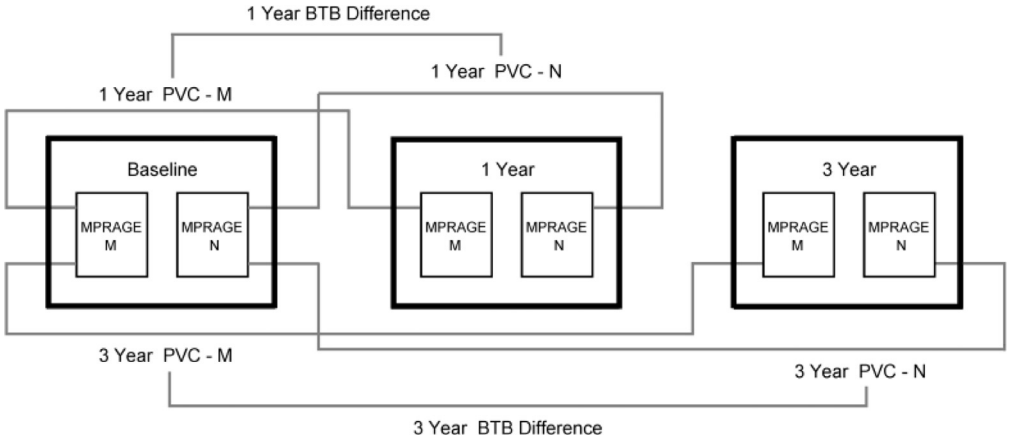

Fig. 1. Calculation of the 1 year and 3 year back to back (BTB) differences for a segmentation method. MPRAGE is the MRI sequence acquired twice at each patient visit with identical parameters (M and N). The percentage volume change (PVC) of each hippocampus is calculated by the segmentation method. The BTB differences over 1 year and 3 years for each hippocampus are calculated by calculating the difference between the respective PVC pairs.

simulated Gaussian distributions with 264 BTB differences each were calculated. After sorting the 10,000,000 ratios, the percentile in the list provides the *p*-value of the null hypothesis. The 95% confidence interval of the ratio was (1.332, 1.665). The 1.3654 ideal ratio is not in the center of the confidence interval because the distribution is not symmetric. The SDBTBD/MDBTBD ratio was also calculated for the 75 subjects with manual segmentations to test the Gaussianity of the manual segmentation. The 10,000,000 simulations were also run and *p*-values calculated for $N = 263$ and 262 BTB differences.

The robust binomial test, which does not depend on Gaussianity or rank (Cover et al., 2016), was used to determine which of any two BTB difference distributions had the larger segmentation noise. The first step of the test calculates the fraction of the first distribution that has a larger absolute BTB difference on a subject by subject basis. In the second step a *p*-value is calculated from the fraction using the binomial distribution and assuming the null hypothesis that two identical distributions will have a fraction of 0.5. The BTB difference distributions were compared for the 1 and 3 year interval for each method and for both the left and right hippocampus. The distributions were also compared against the surrogate manual method, FreeSurfer 5.3.0 longitudinal, to determine which methods had atrophy segmentation noise lower than manual.

A key approach for comparing the performance of two segmentation methods is the number of subjects required to detect a given treatment effect. The relative group size of any two segmentation methods can be calculated from the ratio of the square of the spread of their respective BTB difference distributions. The larger the spread the larger the group size required. For comparison, the relative group size is calculated with two different measures of the spread of the distributions – the MDBTBD and the SDBTBD. For each method the median of the 4 MDBTBD and 4 SDBTBD are used to calculate the relative group sizes.

As manual BTB segmentation of the 264 subjects over 1 and 3 years was unavailable, the segmentation noise of FreeSurfer 5.3.0 in longitudinal mode was used as a surrogate. Previous publications reported that the segmentation noise of FreeSurfer 5.3.0 in fully automatic longitudinal mode is roughly the same as manual segmentation as determined by parametric statistical tests (Mulder et al., 2016) and the similarity of their MDBTBD values (Cover et al., 2016). In the current study the robust binomial test (Cover et al., 2016) is used to verify this result by comparing the BTB difference distributions for the 75 manual segmentation to the FreeSurfer 5.3.0 longitudinal BTB difference distributions for the same subjects.

The analysis of the BTB difference primarily used software routines from Press et al. (2002) with the exception of the Anderson-Darling and Shapiro-Wilk tests which are from the JDistlib library (jdistlib.sourceforge.net).

2.3. Atrophy measurement methods

A total of 5 fully automatic methods were run on each of the ADNI BTB image volumes. Two of the methods were cross sectional – FreeSurfer/ReconAll 5.3.0 in cross sectional mode (Fischl et al., 2012) and FSL/FIRST 5.0.8 (Patenaude et al., 2011) – and 3 were longitudinal – FreeSurfer/ReconAll 5.3.0 in longitudinal mode (Fischl et al., 2012), MAPS-HBSI (Leung et al., 2010) and FreeSurfer/ReconAll 6.0.0 in longitudinal mode (Fischl et al., 2012). All methods were run with their default settings.

It is a common practice to manually review the results of fully automatic segmentation methods and delete any segmentations deemed to be of poor quality. This practice was not performed on any results in the current paper as the goal of the paper is to study performance of the fully automatic methods with no manual assistance.

As an example of an ideal Gaussian distribution, simulated Gaussian BTB difference distributions were generated where each PVC had a segmentation noise with a standard deviation of 2.84% yielding BTB difference distributions with MDBTBDs of roughly 2.2%. This MDBTBD is in line with previous reports of manual segmentation (Mulder et al., 2014, Cover et al., 2016) and fully automatic FreeSurfer 5.3.0 longitudinal – the surrogate for the manual segmentation noise in the current study.

The current study shares some of the aspects of a previous study of atrophy segmentation noise (Cover et al., 2016). The main differences are (1) the inclusion of the 3 year interval for atrophy reproducibility in addition to 1 year, (2) testing the Gaussianity of the atrophy segmentation noise of the hippocampus, (3) assessing the impact of non-Gaussian distributions on statistical calculations such as group size, (4) scatter plots to display the 1 year versus 3 year BTB differences of the segmentation methods, (5) confirming previous reports that fully automatic FreeSurfer 5.3.0 longitudinal has similar noise to manual segmentation with a robust statistical test, and (6) inclusion of the recently released FreeSurfer 6.0.0 segmentation method. For completeness, some of these issues (subjects, study design, MRI acquisition, analysis methods) are discussed in the current paper with the appropriate changes detailed.

3. Results

Table 2 shows a variety of statistics for each of the 4 BTB difference distributions for each of the 7 segmentation methods - which includes the 5 fully automatic segmentation methods, the simulated Gaussian method and the manual method. Examination of the MDBTBDs for each method shows all the MDBTBD for FreeSurfer 6.0.0 longitudinal and MAPS-HBSI are less than for the corresponding values for FreeSurfer 5.3.0 longitudinal – the surrogate for manual segmentation noise. These results indicate MAPS-HBSI and FreeSurfer 6.0.0 longitudinal have

Table 2
Statistics of each of the fully automatic atrophy segmentation methods over 1 and 3 year intervals as well as manual over a 1 year interval.

Method	Left-Right	Years	Number of Patients Yielding BTB Differences	Median Absolute BTB Difference (MDBTBD) [Interquartiles in brackets]	Mean Absolute BTB Difference (MNBTBD)	Standard Deviation Absolute BTB Difference (SDBTBD)	SDBTBD / MDBTBD	p-value Gaussianity Tests			Ratio 1Y = 3Y Binomial Test	*p*-value 1Y = 3Y Binomial Test	Maximum Absolute BTB Difference	Minimum Absolute BTB Diference
								Tailored	Anderson Darling	Shapiro Wilk				
FreeSurfer 5.3.0 Cross Sectional	Left	1	264	3.08 (1.65, 5.56)	4.86	8.51	2.76	0.0000	0.0000	0.0000	0.538	0.1211	62.70	0.02
		3	264	3.10 (1.39, 5.04)	4.34	7.18	2.32	0.0000	0.0000	0.0000			62.34	0.02
	Right	1	264	2.55 (1.05, 4.71)	3.55	5.49	2.15	0.0000	0.0000	0.0000	0.470	0.1477	47.84	0.00
		3	264	3.04 (1.24, 5.40)	4.31	6.86	2.26	0.0000	0.0000	0.0000			54.00	0.00
FreeSurfer 5.3.0 Longitudinal	Left	1	264	1.98 (0.86, 3.31)	2.72	4.01	2.02	0.0000	0.0000	0.0000	0.492	0.3792	20.44	0.01
		3	264	2.21 (0.92, 3.96)	2.95	4.20	1.90	0.0000	0.0000	0.0000			22.98	0.01
	Right	1	264	1.78 (0.85, 3.14)	2.29	2.97	1.67	0.4700	0.2007	0.0892	0.466	0.3792	10.90	0.01
		3	264	2.04 (0.90, 3.73)	2.75	3.75	1.84	0.0006	0.0001	0.0000			14.47	0.01
FreeSurfer 6.0.0 Longitudional	Left	1	264	1.51 (0.68, 2.56)	1.89	2.50	1.66	0.0296	0.0124	0.0044	0.443	0.0281	9.65	0.02
		3	264	1.93 (0.99, 3.26)	2.26	2.84	1.47	0.4389	0.0917	0.1193			8.75	0.00
	Right	1	264	1.36 (0.59, 2.63)	1.89	2.58	1.90	0.0000	0.0000	0.0000	0.428	0.0081	9.91	0.00
		3	264	1.71 (0.92, 2.95)	2.21	2.92	1.71	0.0087	0.0043	0.0003			13.62	0.01
MAPS-HBSI	Left	1	262	1.39 (0.65, 2.42)	1.79	2.42	1.74	0.0086	0.0002	0.0000	0.536	0.1320	9.58	0.02
		3	262	1.31 (0.66, 2.48)	3.55	2.18	1.66	0.0500	0.1772	0.2270			7.91	0.01
	Right	1	262	1.18 (0.59, 2.13)	1.67	2.45	2.07	0.0000	0.0000	0.0000	0.475	0.1931	14.40	0.01
		3	262	1.21 (0.58, 2,36)	3.35	2.34	1.93	0.0000	0.0001	0.0000			10.42	0.01
FSL/FIRST 5.0.8	Left	1	263	2.26 (1.13, 4.15)	3.88	7.70	3.41	0.0000	0.0000	0.0000	0.584	0.0039	54.18	0.01
		3	263	1.73 (0.87, 3.36)	3.55	9.82	5.69	0.0000	0.0000	0.0000			125.62	0.06
	Right	1	263	2.50 (1.27, 5.05)	4.25	7.50	3.00	0.0000	0.0000	0.0000	0.584	0.0039	52.17	0.06
		3	263	1.96 (0.77, 3.94)	3.35	9.05	4.63	0.0000	0.0000	0.0000			128.35	0.00
Simulated Gaussian	Left	1	264	2.21 (1.06, 3.70)	2.62	3.28	1.48	0.4802	0.8831	0.8271	0.508	0.4268	9.08	0.01
		3	264	2.26 (1.07, 3.90)	2.67	3.32	1.47	0.4466	0.8964	0.7566			8.91	0.01
	Right	1	264	1.98 (0.86, 3.62)	2.37	3.02	1.52	0.3029	0.3949	0.8469	0.462	0.0981	9.18	0.01
		3	264	2.16 (0.97, 3.63)	2.51	3.14	1.46	0.4013	0.9334	0.6215			7.77	0.00
Manual	Left	1	75	2.62 (1.02, 4.75)	3.54	5.26	2.01	0.0192	0.0003	0.0000	N/A	N/A	22.34	0.26
	Right	1	75	2.60 (1.01, 4.42)	4.79	16.38	6.31	0.0000	0.0000	0.0000			138.36	0.04

lower segmentation noise than manual. Additionally, all of the MDBTBDs for FreeSurfer 5.3.0 cross sectional and at least 3 of the 4 for FSL/FIRST are larger than the surrogate for manual segmentation. The larger values indicate FreeSurfer 5.3.0 cross sectional and FSL/FIRST are unlikely to be less noisy than manual segmentation. However, detailed knowledge of the distributions of the BTB difference, such as whether they are Gaussian, is needed to assign statistical significance to these MDBTBD comparisons.

Table 2 also shows the *p*-value of the SDBTBD/MDBTBD ratio test of Gaussianity for each BTB difference distribution along with the Anderson-Darling and the Shapiro-Wilk tests. For all 6 the segmentation methods in Table 2, including the manual segmentation, the null hypothesis of a Gaussian distribution is rejected for all three tests. As expected, the simulated Gaussian method did not reject the null hypothesis. Three of the fully automatic methods each have a single BTB difference distribution that does not reject Gaussianity. Using the tailored test, the exceptions are $p = 0.4700$ for FreeSurfer 5.3.0 right hippocampus over one year, $p = 0.4389$ for FreeSurfer 6.0.0 longitudinal left hippocampus over 3 years and $p = 0.0500$ for MAPS-HBSI left hippocampus over 3 years. Nevertheless, none of these 3 distributions should be assumed to be Gaussian as many other properties need to be confirmed than just the SDBTBD/MDBTBD ratio.

The BTB difference distributions of the 5 fully automatic segmentation methods are displayed as scatter plots of 1 year versus 3 years for the left and right hippocampus (Figs. 2 and 3). The manual method is not included as segmentations are only available over a 1 year interval. Review of the scatter plots provides some quick comparison of the segmentation methods independent of any particular statistical test. Scatter plots with the largest clusters correspond to the noisiest segmentation methods. From the scatter plots the methods with the largest noise are FreeSurfer 5.3.0 cross sectional and FSL/FIRST. The methods with the smallest noise are MAPS-HBSI and FreeSurfer 6.0.0 longitudinal. FreeSurfer 5.3.0 longitudinal – the surrogate for manual segmentation – has segmentation noise at the midpoint of the 5 fully automatic methods. Therefore, from the scatter plots, only MAPS-HBSI and FreeSurfer 6.0.0 longitudinal have segmentation noise less than manual. Thus the scatter plots yield an ordering of the methods by segmentation noise consistent with the MDBTBD values.

As the BTB difference distributions are not Gaussian, a binomial test (Cover et al., 2016) – which does not assume a Gaussian distribution and does not depend on ranks - was used to determine the statistical significance of the ordering of the fully automatic methods by segmentation noise. The segmentation noise of the methods was compared to FreeSurfer 5.3.0 longitudinal – the surrogate for the manual segmentation noise. Table 3 presents the fraction and *p*-values of the binomial test for each of the 4 BTB difference distributions for each automatic method. MAPS-HBSI has statistically significant lower segmentation noise than the manual surrogate for all 4 distributions while FreeSurfer 6.0.0 longitudinal has lower noise in 3 of the 4 distributions. The exception for FreeSurfer 6.0.0 longitudinal is $p = 0.0618$ for the 3 year left hippocampus. Thus the binomial test's ordering of the methods by segmentation noise is consistent with the ordering indicated by both the MDBTBD values and the scatter plots.

Table 2 provides the results of the comparison by the binomial test of the segmentation noise of 1 year versus 3 years for each method. MAPS-HBSI has similar segmentation noise over 1 and 3 years as does FreeSurfer 5.3.0 both in cross sectional and longitudinal mode. `

To confirm the role of FreeSurfer 5.3.0 longitudinal as a surrogate for manual segmentation noise, the binomial test was used to compare the BTB difference distributions of manual segmentations to FreeSufer 5.3.0 longitudinal for 75 subjects. For the left hippocampus the fraction was 0.600 and the *p*-value as 0.0527 and for the right it was 0.507 and 0.5000. Thus there was no statistically significant difference between the segmentation noise for manual segmentation and FreeSurfer 5.3.0 longitudinal. Thus FreeSurfer 5.3.0 longitudinal is a valid surrogate for manual segmentation noise.

Table 3 also shows the relative group sizes to detect a specified treatment effect for the 5 fully automatic methods relative to the surrogate for the surrogate manual segmentation noise. The group sizes for both MDBTBD and SDBTBD are provided and, if the Gaussianity assumption held, the two group sizes would be the same. *The biggest discrepancy in group size is for FSL/FIRST that has a relative group size of 1.10 for MDBTBD and 4.66 for SDBTBD* indicating a major break with the Gaussian assumption. The other methods have smaller discrepancies.

4. Discussion

4.1. Gaussianity of the segmentation noise

The most surprising result of the current study is the deviation from Gaussian of most of the BTB difference distributions of all the fully automatic methods as well as the manual method. These results are in contrast to the widespread practice in the literature of using parametric statistical test for segmentation noise studies. Parametric statistical tests should be used for atrophy segmentation noise only when the Gaussianity assumption can be confirmed for a particular distribution.

4.2. Longitudinal segmentation noise over both 1 and 3 years

To date, most of the atrophy reproducibility studies have been over intervals of 1 year or less (Morey et al., 2010; Jovicich et al., 2013; Ramirez et al., 2013; Mulder et al., 2014; Marizzoni et al., 2015; Cover et al., 2016). With multiple year studies becoming more common, knowledge of the behavior of the segmentation noise over several years is becoming more important. The design of the ADNI1 study acquired BTB images at both 1 year and 3 years after baseline for many subjects making it ideal for comparing the segmentation noise over both 1 year and 3 years.

The MDBTBD values, scatter plots and binomial statistical test all agree. MAPS-HBSI and FreeSurfer 6.0.0 longitudinal have substantially lower segmentation noise than the surrogate for manual segmentation. From Table 3, and using the MDBTBD values, MAPS-HBSI only requires 39% of the subjects as manual to detect the same treatment effect. FreeSurfer 6.0.0 longitudinal only requires 49% of the subjects. In contrast both FSL/FIRST and FreeSurfer 5.3.0 cross sectional would require more subjects than manual.

For both MAPS-HBSI and FreeSurfer 6.0.0 longitudinal there was little difference in the segmentation noise over 1 year and 3 years. Therefore, if the atrophy over 3 years was 3 times larger than over 1 year the signal to noise ratio would be 3 times higher over 3 years as well. Thus, as expected, a 3 years interval can be 3 times as sensitive to a treatment effect as a 1 year interval.

It is important to keep in mind the BTB differences are the difference of two atrophy measurements and thus the noise is larger than for a single atrophy measurement. For example, if a atrophy segmentation method with Gaussian noise had a standard deviation of its BTB differences of 2.000 then it would be safe to conclude the standard deviation of the atrophy segmentation noise of the method was 1.414 (= 2/sqrt(2)). However, all the segmentation methods included in the current study have large shoulders compared to a Gaussian distribution and are therefore not Gaussian. Moreover, the large shoulders yield outlying atrophy values. When BTB difference of two atrophies is calculated the chances of an outlying point roughly doubles because either of the atrophies could yield one. Thus the 1 year versus 3 years BTB difference scatter plots in Figs. 2 and 3 show many outlying points with the exception of the simulated Gaussian distribution.

4.3. Group size calculations

Group size calculations are also affected by the lack of Gaussianity of the segmentation noise distributions. Since the ratio of MDBTBD/

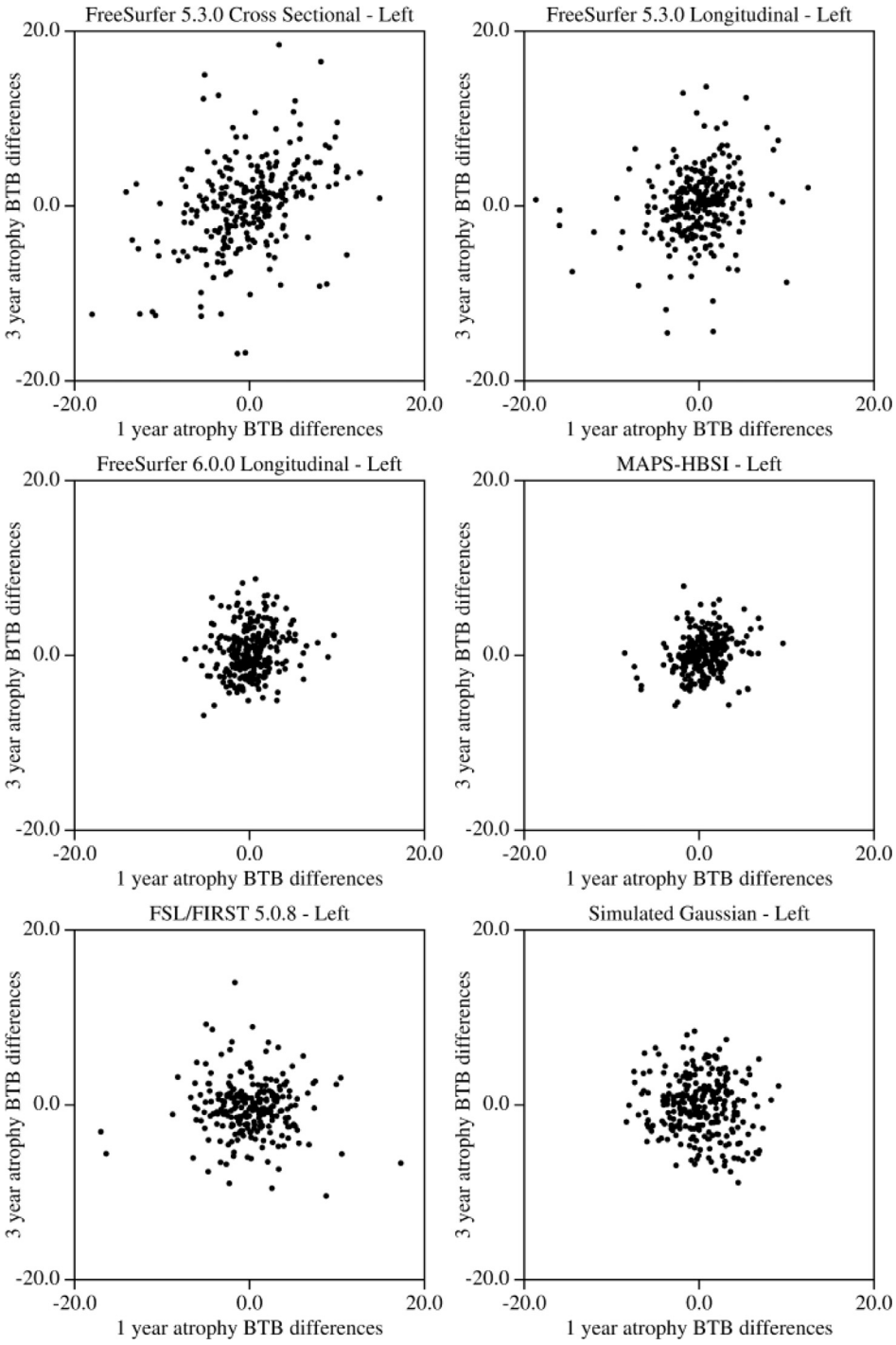

Fig. 2. Scatter plots for the left hippocampus of the 6 segmentation methods for the BTB differences of 1 year versus 3 years. The smaller the clusters the smaller the noise.

SDBTBD for a Gaussian distribution is always 1.3654, if the BTB difference distributions were all Gaussian, the calculations would yield the same group size whether the square of the MDBTBD or SDBTBD was used to calculate the group size. However, as was established in the Results section, the distributions are not Gaussian and thus the relative group sizes may differ between MDBTBD and SDBTBD. As the SDBTBD is particularly sensitive to the few BTB differences in the large shoulders it is recommended to use the MDBTBD rather than SDBTBD to calculate group sizes as group size calculation are than less likely to be influenced by a few subjects with outlying points.

4.4. Calculating p-values from BTBD distributions

Several different types of statistical tests have been used in the literature to calculate a *p*-value when comparing two BTB difference distributions. All statistical tests require the BTB differences for the two methods being compared. These tests include parametric tests which assume a Gaussian distribution, rank tests (Marizzoni et al., 2015; Cash et al., 2015) such at the Wilcoxon paired signed test and a test based on the binomial distribution (Cover et al., 2016). However, parametric tests should be avoided unless the Gaussianity the distributions is confirmed.

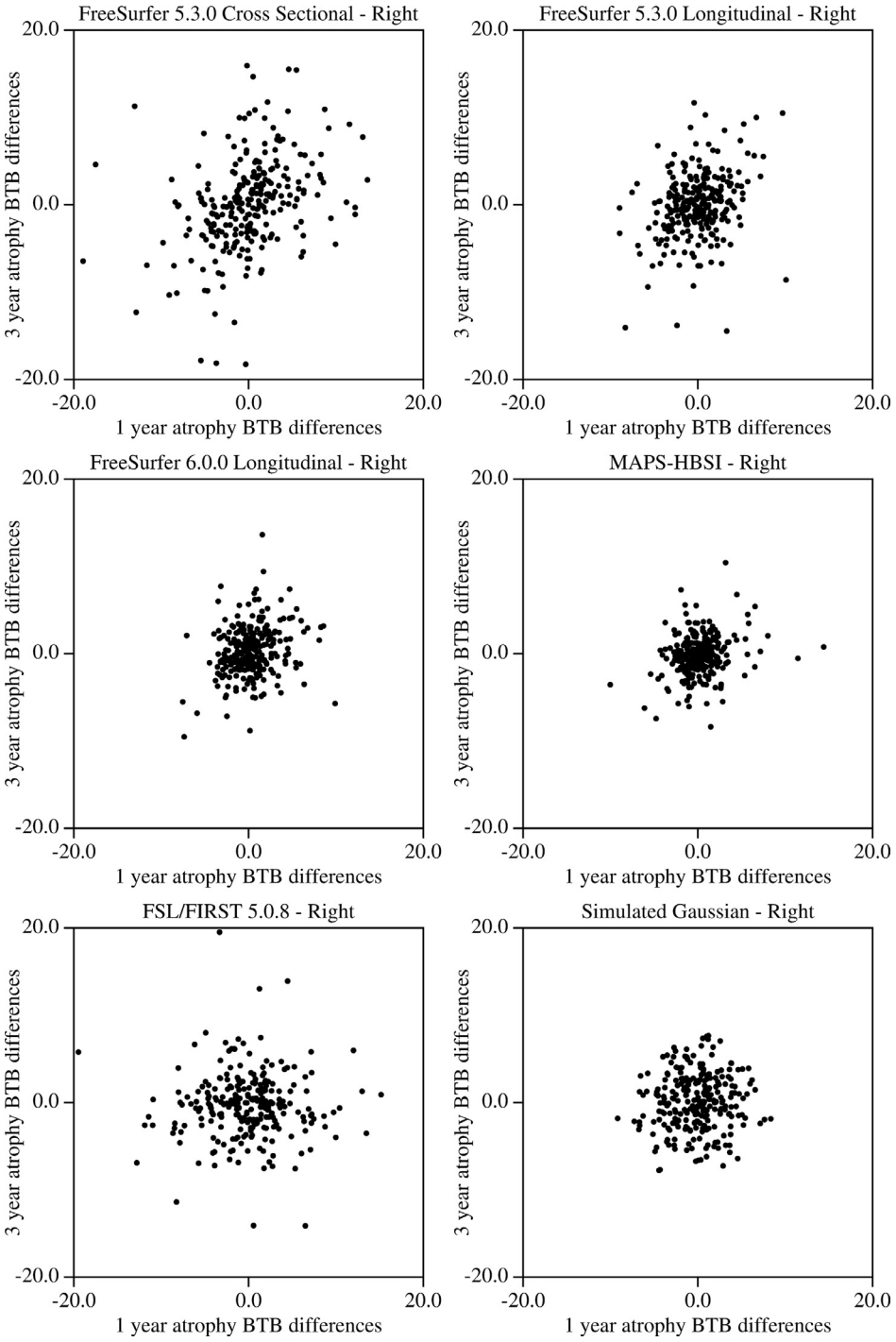

Fig. 3. Scatter plots for the right hippocampus of the 6 segmentation methods for the BTB differences of 1 year versus 3 years. The smaller the clusters the smaller the noise.

The binomial test makes the least assumptions regarding the BTB difference distributions and are therefore most likely to be valid. The Wilcoxon test includes the assumption that all subjects are measured with the same scale. For example, the Wilcoxon test assumes, that for a given method, that a 1% BTB difference is comparable for two subjects - one with an annual change of 0% and another with an annual change of 5%. Given the complicated algorithms used in the some of the segmentation methods released in recent years, it is unclear if this comparison is valid for all methods. As the binomial test only compares values within each subject it avoids this assumption. It is important to keep in mind, that for any two methods being compared, the median BTB difference should be the same, ensuring the two methods have the same scale.

A major advantage of using statistical tests with more assumptions is, provided the assumptions hold, a significant result can be found with fewer subjects. In clinical trials inclusion of subjects is often expensive and analysis of the data for each subject may also be expensive. Therefore, using a statistical test with more assumptions can substantially reduce the number of subjects and thus the cost of a clinical trial. Consequently, the use of statistical tests with a large number of assumptions is often justified. However, if the assumptions of a statistical test do not hold, such as Gaussianity or rank, then spurious results

Table 3
Comparison of the segmentation methods to the surrogate for the noise of manual segmentation - FreeSurfer 5.3.0 longitudinal. For the binomial comparison, when the binomial fraction is less than 0.5 the method is less noisy than surrogate manual segmentation.

Method	MDBTBD group size relative to FreeSurfer 5.3.0 longitudinal		SDBTBD group size relative to FreeSurfer 5.3.0 longitudinal		Binomial test greater than FreeSurfer 5.3.0 longitudinal (FreeSurfer 5.3.0 longitudinal is a surrogate for manual segmentation noise)							
					Left				Right			
	Median (%)	Relative Size	Median (%)	Relative Size	1 year Fraction	p-value	3 year Fraction	p-value	1 year Fraction	p-value	3 year Fraction	p-value
FreeSurfer 5.3.0 cross sectional	3.06	2.32	7.02	3.27	0.670	0.0000	0.655	0.0000	0.595	0.0012	0.655	0.0000
FreeSurfer 5.3.0 longitudinal	2.01	1.00	3.88	1.00	0.500	0.4750	0.500	0.4754	0.500	0.4750	0.500	0.4754
FreeSurfer 6.0.0 longitudinal	1.63	0.64	2.76	0.49	0.402	0.0005	0.398	0.0003	0.382	0.0005	0.454	0.0618
MAPS-HBSI	1.26	0.39	2.38	0.38	0.405	0.0008	0.348	0.0000	0.401	0.0050	0.351	0.0000
FSL/FIRST 5.0.8	2.11	1.10	8.38	4.66	0.536	0.1340	0.423	0.0056	0.635	0.0000	0.519	0.2891

will occur. Thus the statistical test with the least number of assumptions is the safe bet.

Fortunately, the cost of subjects for measuring the segmentation noise of a fully automatic method is essentially zero. The 264 subjects included in the current study were downloaded from the public ADNI web site at no cost. While the fully automatic segmentation of the subject image volumes can take several months on high performance computers the CPU time, disk space and other resources are often low cost or free. As the current study has demonstrated, a cohort of 264 subjects is usually sufficient to clearly differentiate between the performance of segmentation methods. Thus for calculating a *p*-value when comparing the segmentation noise of hippocampal atrophy measures the cost benefits ratio for the use of parametric type or Wilcoxon statistical tests are less relevant and the robust and dependable binomial test is the prudent choice.

In the current paper we chose to use a novel technique for testing if a distribution is Gaussian tailored to outlying points in addition to two conventional tests. The technique was designed by finding the ratio of two parameters from the segmentation literature, SDBTBD and MDBTBD, and thus was tailored to answered one of the primary questions of the current paper. While there are a host of different techniques presented in the literature for testing whether a sampled distribution is Gaussian - including the D'Agostino's K-squared test, Jarque–Bera test, Anderson–Darling test, Cramér–von Mises criterion, Lilliefors test, Kolmogorov–Smirnov test, Shapiro–Wilk test, Pearson's chi-squared test, skewness and kurtosis (Razali and Wah, 2011), none of these tests are as well suited to the problem of outlying points as the tailored test as we know the departure from Gaussianity was due to outlying points and not some other deviation from Gaussianity. For example, there is no indications the skewness of the BTBD distributions is causing major problems with the statistics.

4.5. Future work

A clinical key question in the field of hippocampal atrophy segmentation is whether the noise of current segmentation methods can be reduced so segmentation can more often aid in the diagnosis and treatment of individual subjects. From the ADNI1 study the median annualized atrophy rates were 1.5% (HC), 2.4% (MCI) and 5.1% (AD) (Cover et al., 2016). The segmentation method in the current study with the lowest median MDBTBD, MAPS-HBSI, had a value of 1.26%. From these values, the difference between the atrophy over 1 year of HC and MCI is 0.9%. If there was a segmentation method with a MDBTBD of 0.13%, 1/10 of the current MAPS-HBSI value, it may be possible to determine if a subject has advanced to MCI just based on an atrophy measurement. Could a hippocampal segmentation method with such low noise be possible in the next future?

A hint is provided by the comparison of the segmentation noise of MAPS-HBSI at 1.5T and 3T (Cover et al., 2016). Using the robust binomial statistical test, it was found the segmentation noise at 1.5T and 3T were the same in spite of the fact that 3T scanners have lower instrumentation noise. This result suggests that currently the segmentation noise is dominating the MRI scanner's instrumentation noise. Therefore, it may be possible to improve segmentation methods to where MRI's noise is the source of the floor on the segmentation noise. However, until there is a working method with a low enough segmentation noise to be sensitive to the lower instrumentation noise of a 3T MRI it is impossible to be certain such a segmentation method is possible.

One of the goals of the current paper is to provide a standard against which the noise of new segmentation methods can be compared. As all of the ADNI1 images files used in the current paper are listed in Table S1, it is possible to reproduce all the fully automatic calculations in the current paper as well as the 1 year versus 3 years scatter plots of the BTB differences. Therefore, it is also possible to run the same analysis on new algorithms and see if the segmentation noise is better than the segmentation methods presented in the current paper.

4.6. Conclusions

Fully automatic FreeSurfer 6.0.0 and MAPS-HBSI both have lower segmentation noise than manual requiring 64% to 39% of the subjects, respectively, to detect the same treatment effect as manual. Fully automatic FreeSurfer 5.3.0 cross sectional and FSL/FIRST 5.0.8 have hippocampal segmentation noise no better than manual. Fully automatic FreeSurfer 5.3.0 longitudinal has similar noise to manual segmentation for the hippocampus and can be used as a fully automatic surrogate for manual segmentation noise.

All the methods evaluated had segmentation noise distributions which violated the Gaussianity assumption. Therefore, robust statistics, such as the MDBTBD and the binomial test, should be used to summarize the noise of segmentation methods. Given the diverse nature of the BTB difference distributions of segmentation noise, Gaussianity should be confirmed before the use of parametric statistics as should any other statistical assumptions underlying statistical tests. Studies of new or improved hippocampal segmentation methods should employ a core standard set of statistical tests of the segmentation noise so that the performance of the methods can be compared to those in the literature.

Funding Information

The “neuGRID4you” project (2011-2015) was funded from the European Commission's Seventh Framework Programme (FP7/2007–2013) under grant agreement no.283562.

Software and data used in this paper were partially a product of the "neuGRID4you" project.

None of the authors have a conflict of interest.

Acknowledgements

Study funding was provided by neuGRID4you (www.neuGRID4you.eu), a European Community FP7 project (grant agreement 283562), and the VU University Medical Center, Amsterdam, The Netherlands. All calculations were performed on the NCAGRID cluster at the VU University Medical Center. Thanks to Wessl N Wieringen of VUmc's Epidemiology and Biostatistics Department for checking the statistical analysis.

Data collection and sharing for this 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). 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 organization 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 Neuro Imaging at the University of Southern California.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.pscychresns.2018.06.011.

References

Ahdidan, J., Raji, C.A., DeYoe, E.A., Mathis, J., Noe, K.Ø., Rimestad, J., et al., 2015. Quantitative neuroimaging software for clinical assessment of hippocampal volumes on MR imaging. J. Alzheimers Dis. 49, 723–732.

Bobinski, M., Wegiel, J., Wisniewski, H.M., Tarnawski, M., Bobinski, M., Reisberg, B., et al., 1996. Neurofibrillary pathology–correlation with hippocampal formation atrophy in Alzheimer disease. Neurobiol. Aging. 17, 909–919.

Cash, D.M., Frost, C., Iheme, L.O., Ünay, D., Kandemir, M., Fripp, J., et al., 2015. Assessing atrophy measurement techniques in dementia: Results from the MIRIAD atrophy challenge. Neuroimage 123, 149–164.

Chincarini, A., Sensi, F., Rei, L., Gemme, G., Squarcia, S., Longo, R., et al., 2016. Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease. Neuroimage 125, 834–847.

Cover, K.S., van Schijndel, R.A., van Dijk, B.W., Redolfi, A., Knol, D.L., Frisoni, G.B., et al., 2011. Assessing the reproducibility of the SienaX and Siena brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. Psychiatry Res. 193, 182–190.

Cover, K.S., van Schijndel, R.A., Versteeg, A., Leung, K.K., Mulder, E.R., Jong, R.A., et al., 2016. Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease. Psychiatry Res. 252, 26–35.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774–781.

Frisoni, G.B., Fox, N.C., Jack Jr, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6, 67–77.

Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., et al., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9, 119–128.

Jack Jr, C.R., Vemuri, P., Wiste, H.J., Weigand, S.D., Aisen, P.S., Trojanowski, J.Q., et al., 2011. Evidence for ordering of Alzheimer disease biomarkers. Arch. Neurol. 68, 1526–1535.

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartrés-Faz, D., Arnold, J., et al., 2013. Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. Neuroimage 83, 472–484.

Marizzoni, M., Antelmi, L., Bosch, B., Bartrés-Faz, D., Müller, B.W., Wiltfang, J., et al., 2015. Longitudinal reproducibility of automatically segmented hippocampal subfields: A multisite European 3T study on healthy elderly. Hum. Brain Mapp. 36, 3516–3527.

Morey, R.A., Selgrade, E.S., Wagner 2nd, H.R., Huettel, S, Wang, L., McCarthy, G., 2010. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Hum. Brain Mapp. 31, 1751–1762.

Mulder, E.R., de Jong, R.A., Knol, D.L., van Schijndel, R.A., Cover, K.S., Visser, P.J., Barkhof, F., Vrenken, H. Hippocampal volume change measurement: quantitive assessment of the reproducibility of expert manual outling and the automated methods FreeSurfer and FIRST. LEAP. Available at: http://www.ixico.com/additional-information/leap-analysis. Accessed December 14, 2014.

NeuroQuant. Available at: https://www.advancedradiology.com/our-services/mri/neuroquant%C2%AE. Accessed December 14, 2016.

NeuroReader. Available at: http://brainreader.net/p/product/neuroreader. Accessed December 14, 2016.

Ochs, A.L., Ross, D.E., Zannoni, M.D., Abildskov, T.J., Bigler, E.D., Alzheimer's Disease Neuroimaging Initiative, 2015. Comparison of automated brain volume measures obtained with NeuroQuant and FreeSurfer. J. Neuroimaging 25, 721–727.

Opfer, R., Suppa, P., Kepp, T., Spies, L., Schippling, S., Huppertz, H.J., et al., 2016. Atlas based brain volumetry: How to distinguish regional volume changes due to biological or physiological effects from inherent noise of the methodology. Magn. Reson. Imaging 34, 455–461.

Patenaude, B., Smith, S.M., Kennedy, D., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain. NeuroImage 56, 907–922.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2002. Numerical Recipes in C; The art of scientific computing. Cambridge University Press, Cambridge.

Ramirez, J., Scott, C.J., Black, S.E., 2013. A short-term scan-rescan reliability test measuring brain tissue and subcortical hyperintensity volumetrics obtained using the lesion explorer structural MRI processing pipeline. Brain Topogr. 26, 35–38.

Razali, N., Wah, Y.B., 2011. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. J. Stat. Model. Analyt. 2, 21–33.

Smith, S.M., Rao, A., De Stefano, N., Jenkinson, M., Schott, J.M., Matthews, P.M., et al., 2007. Longitudinal and cross-sectional analysis of atrophy in Alzheimer's disease: cross-validation of BSI, SIENA and SIENAX. Neuroimage 36, 1200–1206.

Wolz, R., Schwarz, A.J., Yu, P., Cole, P.E., Rueckert, D., Jack Jr, C.R., et al., 2014. Robustness of automated hippocampal volumetry across magnetic resonance field strengths and repeat images. Alzheimers Dement 10, 430–438.