

Reproducibility of Deep Gray Matter Atrophy Rate Measurement in a Large Multicenter Dataset

 A. Meijerman,  H. Amiri,  M.D. Steenwijk,  M.A. Jonker,  R.A. van Schijndel,  K.S. Cover, and  H. Vrenken, for the Alzheimer's Disease Neuroimaging Initiative



ABSTRACT

BACKGROUND AND PURPOSE: Precise in vivo measurement of deep GM volume change is a highly demanded prerequisite for an adequate evaluation of disease progression and new treatments. However, quantitative data on the reproducibility of deep GM structure volumetry are not yet available. In this paper we aim to investigate this reproducibility using a large multicenter dataset.

MATERIALS AND METHODS: We have assessed the reproducibility of 2 automated segmentation software packages (FreeSurfer and the FMRIB Integrated Registration and Segmentation Tool) by quantifying the volume changes of deep GM structures by using back-to-back MR imaging scans from the Alzheimer Disease Neuroimaging Initiative's multicenter dataset. Five hundred sixty-two subjects with scans at baseline and 1 year were included. Reproducibility was investigated in the bilateral caudate nucleus, putamen, amygdala, globus pallidus, and thalamus by carrying out descriptives as well as multilevel and variance component analysis.

RESULTS: Median absolute back-to-back differences varied between GM structures, ranging from 59.6–156.4 μL for volume change, and 1.26%–8.63% for percentage volume change. FreeSurfer had a better performance for the outcome of longitudinal volume change for the bilateral amygdala, putamen, left caudate nucleus ($P < .005$), and right thalamus ($P < .001$). For longitudinal percentage volume change, FreeSurfer performed better for the left amygdala, bilateral caudate nucleus, and left putamen ($P < .001$). Smaller limits of agreement were found for FreeSurfer for both outcomes for all GM structures except the globus pallidus. Our results showed that back-to-back differences in 1-year percentage volume change were approximately 1.5–3.5 times larger than the mean measured 1-year volume change of those structures.

CONCLUSIONS: Longitudinal deep GM atrophy measures should be interpreted with caution. Furthermore, deep GM atrophy measurement techniques require substantially improved reproducibility, specifically when aiming for personalized medicine.

ABBREVIATIONS: AD = Alzheimer disease; ADNI = Alzheimer Disease Neuroimaging Initiative; BTB = back-to-back; FIRST = FMRIB Integrated Registration and Segmentation Tool; LoA = limit of agreement; MCI = mild cognitive impairment; SEM = standard error of measurement

Neurodegeneration occurs in Alzheimer disease (AD). The process is characterized by neuronal loss and axonal and

synaptic degeneration.^{1–4} Growing evidence reveals that this process happens within early phases of the disease and before making

Received June 16, 2017; accepted after revision August 28.

From the Departments of Radiology and Nuclear Medicine (A.M., H.A., M.D.S., R.A.v.S., K.S.C., H.V.) and Epidemiology and Biostatistics (A.M., M.A.J.), Vrije University Medical Center, Amsterdam, The Netherlands; and the Neuroscience Research Center, Institute of Neuropharmacology (H.A.), Kerman University of Medical Sciences, Kerman, Iran.


This work was supported by the Alzheimer Disease Neuroimaging Initiative (ADNI) (National Institutes of Health grant U01 AG024904) and Department of Defense 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.; 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

A. Meijerman and H. Amiri contributed equally to this work.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data base (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 the 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.

Please address correspondence to Houshang Amiri, PhD, Vrije University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam; e-mail: amiri.houshang@gmail.com

 Indicates open access to non-subscribers at www.ajnr.org

<http://dx.doi.org/10.3174/ajnr.A5459>

a clinical diagnosis.^{5,6} The development of neurodegeneration on a large scale during disease leads to loss of tissue volume (the so-called atrophy), which can be quantified by using structural MR imaging.

Atrophy has been found to be associated with impaired neurologic and neurocognitive performance.⁷⁻¹⁰ More recently, research revealed that deep GM atrophy specifically plays an important role in the characterization, course, and progression of AD¹¹⁻¹⁷ and in other diseases like MS¹⁸⁻²⁰ and Parkinson disease.²¹⁻²³ Measurements of deep GM atrophy could therefore be of importance in the evaluation of neuroprotective treatment (eg, in investigating drug efficacy). Currently, a growing number of clinical trials are incorporating brain volume changes as an early biomarker.²⁴ To use atrophy as a reliable biomarker for the extent of neurodegeneration and axonal damage, the precision and reproducibility of volume change measurement techniques should be evaluated. Of note, having precise and reproducible methods would increase statistical power, which reduces sample sizes for detecting effects in clinical trials.

Among automated tissue segmentation software for deep GM structures, FreeSurfer (<http://surfer.nmr.mgh.harvard.edu>)²⁵ and the FMRIB Integrated Registration and Segmentation Tool (FIRST; part of FSL, <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST>)²⁶⁻²⁹ are freely available and widely used. Whereas FreeSurfer has a longitudinal pipeline by which multiple time points can be analyzed, FIRST is a cross-sectional technique that analyses only a single time point. Despite the importance of the measurement of deep GM atrophy rate, little is known about reproducibility of the measurements over time in large multicenter datasets.

In this paper, to assess reproducibility, we used data from the Alzheimer Disease Neuroimaging Initiative (ADNI) study³⁰ acquired at 1.5T, including 2 back-to-back (BTB) 3D T1-weighted images at each time point.³¹ We quantified reproducibility by using BTB differences of 1-year volume change and of percentage volume change for the bilateral amygdala, caudate nucleus, globus pallidus, putamen, and thalamus. To this end, we used 3 different statistical methods. First, we used descriptive statistics by which median absolute differences are reported. This method is frequently used, but its outcome measures cannot be compared statistically between methods. Therefore, we additionally used analytical statistics based on the difference in the regression coefficient. Lastly, we used the method of determination of the standard error of measurement, which very precisely maps reproducibility by modeling different components related to variability in BTB measures.

MATERIALS AND METHODS

ADNI Dataset

Data used in this study were taken from the ADNI1 study.³⁰ The primary goal of the ADNI has been to test whether serial MR imaging, PET, other biologic markers, and clinical and neuropsychologic assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early AD.

A total of 800 included subjects from 55 sites in the US and Canada were enrolled between 2004 and 2010 and were followed up in a 2- to 3-year time interval. Written informed consent was

obtained before each baseline visit. Inclusion criteria were age between 55–90 years, having a study partner able to provide an independent evaluation of functioning, and speaking either English or Spanish. All subjects were willing and able to undergo all test procedures including neuroimaging and agreed to longitudinal follow-up. Exclusion criteria were specific psychoactive medications. For control subjects, inclusion criteria were as follows: Mini-Mental State Examination scores between 24–30 (inclusive), a clinical dementia rating of 0, and no history of depression, MCI, and dementia. The age range was matched to that of MCI and AD subjects. For subjects with MCI, inclusion criteria were as follows: Mini-Mental State Examination scores between 24–30 (inclusive), a memory complaint, objective memory loss measured by education-adjusted scores on the Wechsler Memory Scale Logical Memory II, a clinical dementia rating of 0.5, absence of high levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. For subjects with mild AD, inclusion criteria were as follows: Mini-Mental State Examination scores between 20–26 (inclusive), clinical dementia rating of 0.5 or 1.0, and meets National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association criteria for probable AD. A standardized imaging protocol carried out over qualified sites included the acquisition of 2 sequential 3D T1-weighted MPRAGE scans (ie, BTB) at baseline and at the 1-year study time point.³²

Subjects

Our study involved 562 subjects who had exactly 2 MPRAGE scans acquired at both the baseline and at 1 year, with 3D T1-weighted BTB images acquired at both time points at 1.5T. Three hundred twenty-two (57.3%) subjects were male and 240 (42.7%) were female. The median age at baseline was 75.3 years (interquartile range, 8.7). One hundred fourteen (30.4%) were diagnosed with probable AD, 277 (49.3%) with MCI, and 171 (20.3%) were healthy controls. Data were requested after written compliance to the ADNI data use agreement and data sharing policy and were obtained from the ADNI data image and data archive LONI (Laboratory of Neuro Imaging; <http://adni.loni.usc.edu>). All data were received anonymized by ADNI procedures and with assignment of a unique ADNI study number to subjects.

Volumetric Measurements

MR image acquisition included standard automated adjustments with no additional postprocessing such as intensity nonuniformity correction or gradient warp correction. DICOM images of subjects were converted to NIFTI format for further processing by using `dicom2nifti` (<http://www.cabiatl.com/mricro/mricron/dcm2nii.html>).

Automated deep GM segmentations were performed on the NCAgrid (a 64-bit Linux computer cluster with 512 cores) by using 2 freely available and frequently used software packages: FreeSurfer version 5.3.0²⁵ and FIRST implemented in FSL version 5.0.8.²⁶⁻²⁹

For FreeSurfer, images were segmented by using the longitudinal image processing stream, which analyzes 2 time points simultaneously to improve the estimation of volumes and volume

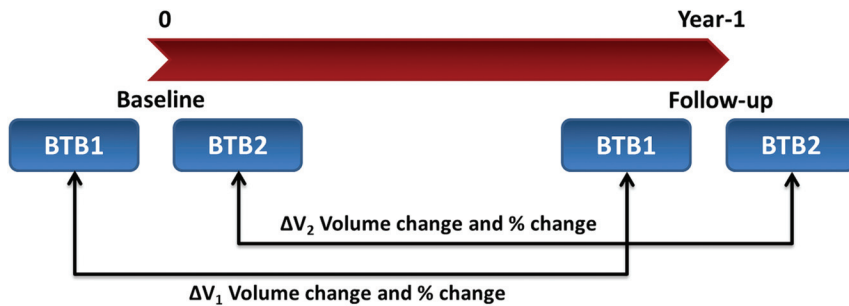


FIG 1. Scheme showing both BTB scans at each time point and calculation of the volume change and percentage volume change.

change. Within FIRST, the default parameters were used.²⁵ Segmentations were carried out for both BTB scans at baseline and at the 1-year study time point, leading to a total number of 134,880 segmentations.

Outcome Measures

The 2 derived main outcome measures in our study were the longitudinal volume change and percentage volume change. The volume change (ΔV , in μL) was calculated for each longitudinal scan pair (two BTB1 and two BTB2) as:

$$\Delta V_1 = V^{\text{Year 1(BTB1)}} - V^{\text{Baseline(BTB1)}}$$

and

$$\Delta V_2 = V^{\text{Year 1(BTB2)}} - V^{\text{Baseline(BTB2)}}$$

The percentage volume change for both ΔV_1 and ΔV_2 was calculated separately as:

$$100 \times (\Delta V \div V^{\text{Baseline}})$$

Fig 1 schematically shows study time points and the calculation of the volume change and percentage volume change by using BTB scans.

In both BTB scans (BTB1 and BTB2), at each time point, the brain is assumed to be identical; therefore $\Delta V_2 - \Delta V_1$ can be used as a measure of reproducibility for each outcome measure (ie, absolute and percentage volume change).

Statistical Analysis

Data distribution and missing data were carefully checked before all statistical analyses. Reproducibility according to BTB scans is reported by using 3 methods of analysis for both software packages. First, we used median absolute BTB differences. Second, we compared the absolute BTB differences based on differences in the regression coefficient (effect size). This involved the construction of separate linear multilevel models for each deep GM structure and each hemisphere. Data were natural log-transformed before analysis to avoid fitting the model to a skewed distribution of our data. In our multilevel models, a random intercept was chosen to correct for the dependency of observations clustering within each same subject. Variance around the intercept was assumed to be normally distributed. Statistics were reported as *P* values, back-transformed effect sizes, and their corresponding 95% confidence intervals.

Finally, as a third method, we assessed reproducibility by determining the limit of agreement (LoA), which is considered as a

very sensitive method of analysis.³³⁻³⁵ This was done by constructing separate linear multilevel models for each deep GM structure summing the variance components attributable to BTB scans to determine the level of random bias in both outcome variables. Because the method is based on variance, contrary to the first 2 methods, it uses the original (nonabsolute) values of each volume change analysis. Fixed factors in our multilevel model included hemisphere, software package (FreeSurfer or FIRST),

sex, diagnostic group, and all possible interactions between these variables. Random factors in the model included hemispheres, software package, time point, all possible interactions between them, and the use of a random intercept on the subject level. Nesting of the factors was carried out according to the method described by Mulder and colleagues.³⁵ We used restricted maximum likelihood as the estimating procedure in all multilevel analyses and assumed an independent covariance matrix. The best fitting model to the data was then chosen based on the lowest Akaike information criterion. Interscan standard errors of measurement (SEMs) attributable to BTB scans for each software package were calculated by summing the random variance components of the multilevel models related to BTB (ie, the variance attributable to the interaction between the random chosen variables and time point; see Equation 1 below). The separate variance components required to sum SEM were assumed to be independent of each other. The variance component containing the highest interaction (ie, σ^2 [time point \times hemisphere \times software package]) was considered to be completely part of the error variance in our calculations. Furthermore, all variance components containing a time point were allowed to vary within software package.

$$1) \text{ SEM}^2 = \sigma^2 (\text{time point}) + \sigma^2 (\text{time point} \times \text{hemisphere}) + \sigma^2 (\text{time point} \times \text{software package}) + \sigma^2 (\text{error})$$

Then, LoA, as a measure of reproducibility, was derived and reported from the SEM for each software package by using Equation 2. The lower the LoA, the better the reproducibility.

$$2) \text{ LoA} = \pm 1.96 \times \sqrt{2} \times \text{SEM}$$

The quality of all MR images was inspected visually. Regarding the quality of the segmentation, we identified severe outliers based on implausible results of the outcome measures. Implausible outliers in terms of longitudinal volume change or percentage longitudinal volume change were considered to be a consequence of a failure in segmentation. An implausible outlier was identified if the longitudinal BTB difference was more than 25% of its corresponding baseline volume. We created separate linear multilevel models with and without implausible large outliers to evaluate their impact on our SEM. These outliers were treated as missing data in our final analysis. In addition, we compared the number of outliers between FreeSurfer and FIRST in all deep GM structures. This was carried out by using the binominal

Table 1: Nonannualized atrophy rates for deep GM structures for each hemisphere per group

GM Structure	Software	Hemisphere	Atrophy Rate for Control Patients, %	Atrophy Rate for Patients with MCI, %	Atrophy Rate for Patients with AD, %
Caudate nucleus	FreeSurfer	Left	-0.33	-0.95	-1.63
	FIRST	Left	-0.84	-0.78	-1.72
	FreeSurfer	Right	-0.35	-0.78	-1.58
	FIRST	Right	-1.07	-0.66	-1.42
Putamen	FreeSurfer	Left	-0.04	-0.69	-2.50
	FIRST	Left	-0.44	-1.20	-2.16
	FreeSurfer	Right	-0.37	-0.73	-1.72
	FIRST	Right	-0.61	-1.11	-1.71
Amygdala	FreeSurfer	Left	-0.56	-2.70	-4.67
	FIRST	Left	-0.90	-1.59	-4.36
	FreeSurfer	Right	-0.70	-2.25	-3.95
	FIRST	Right	-0.38	-3.60	-3.57
Globus pallidus	FreeSurfer	Left	-0.45	0.25	0.58
	FIRST	Left	-0.10	-0.92	-1.43
	FreeSurfer	Right	-0.11	-0.38	-0.50
	FIRST	Right	-0.67	-0.90	-1.87
Thalamus	FreeSurfer	Left	-0.88	-1.69	-2.06
	FIRST	Left	-0.90	-0.85	-1.05
	FreeSurfer	Right	-0.78	-1.38	-2.29
	FIRST	Right	-0.62	-0.71	-0.94

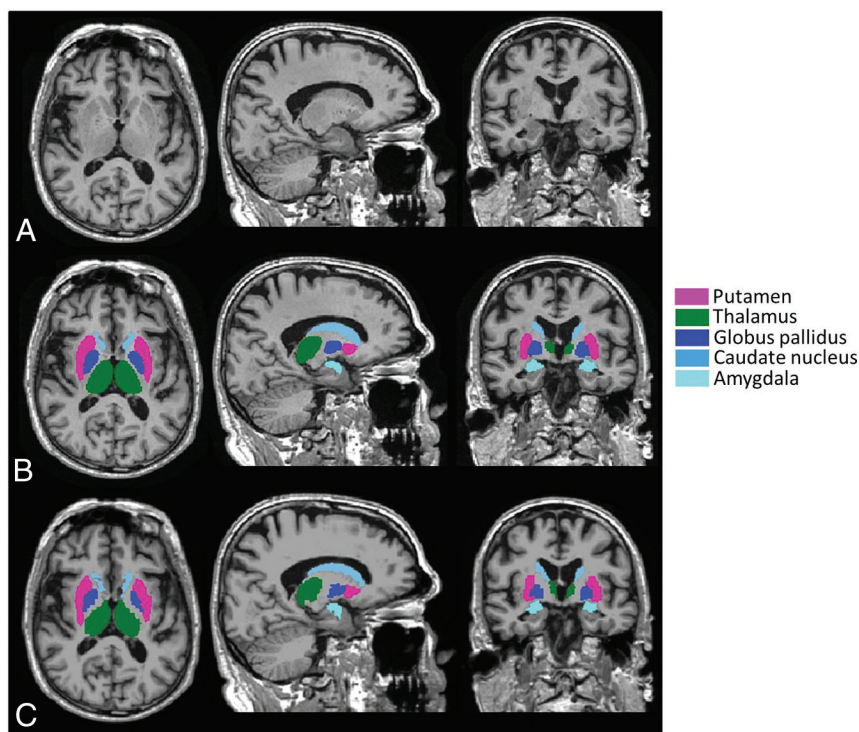


FIG 2. A, An example of a 3D T1-weighted image segmented with B, FIRST and C, FreeSurfer.

test, which tested an equal distribution of the number of outliers for both FreeSurfer and FIRST.

For illustrating agreement, Bland-Altman plots were created. A Bland-Altman plot represents the difference in BTB of an outcome measure versus its mean.^{36,37} We created plots for both outcome measures of FreeSurfer and FIRST, with and without implausible outliers. In this paper, for this method, we present the results of analysis performed on data excluding implausible outliers.

All statistical analysis was carried out by using SPSS version 21 (IBM, Armonk, New York) except for the modeling of data to obtain SEM and derived LoAs, which was carried out by using SAS Studio

version 3.4 (SAS Institute, Cary, North Carolina). The level of significance in our models was set to 0.05 (5%).

RESULTS

Median follow-up time ($\chi^2 = 1.566$; *df*, 2; $P = .45$) and age ($\chi^2 = 0.992$; *df*, 2; $P = .60$) did not differ between the 3 study groups. To enable a direct comparison of reproducibility metrics to the measured (percentage) volume change values, nonannualized median atrophy rates are presented in Table 1. As expected, atrophy rates were generally higher in patients with AD compared with patients with MCI and control patients, with the highest rates found for the amygdala. For 2 different male healthy control patients, FreeSurfer and FIRST segmentation failed. Therefore, for each software package, 561 subjects were included in the longitudinal data analysis. A typical example of FreeSurfer and FIRST segmentations is shown in Fig 2. BTB differences are illustrated by the example in Fig 3, which shows Bland-Altman plots of BTB difference in longitudinal volume change for the left caudate nucleus for both FreeSurfer and FIRST, excluding the improbable outliers.

Descriptive Statistics

Descriptive statistics for each hemisphere for each deep GM structure for measuring longitudinal volume change and longitudinal percentage volume change are presented in Tables 2 and 3, respectively. Based on these reported descriptive statistics (median absolute BTB differences with corresponding 90th percentile indicating spread), as expected, the smaller

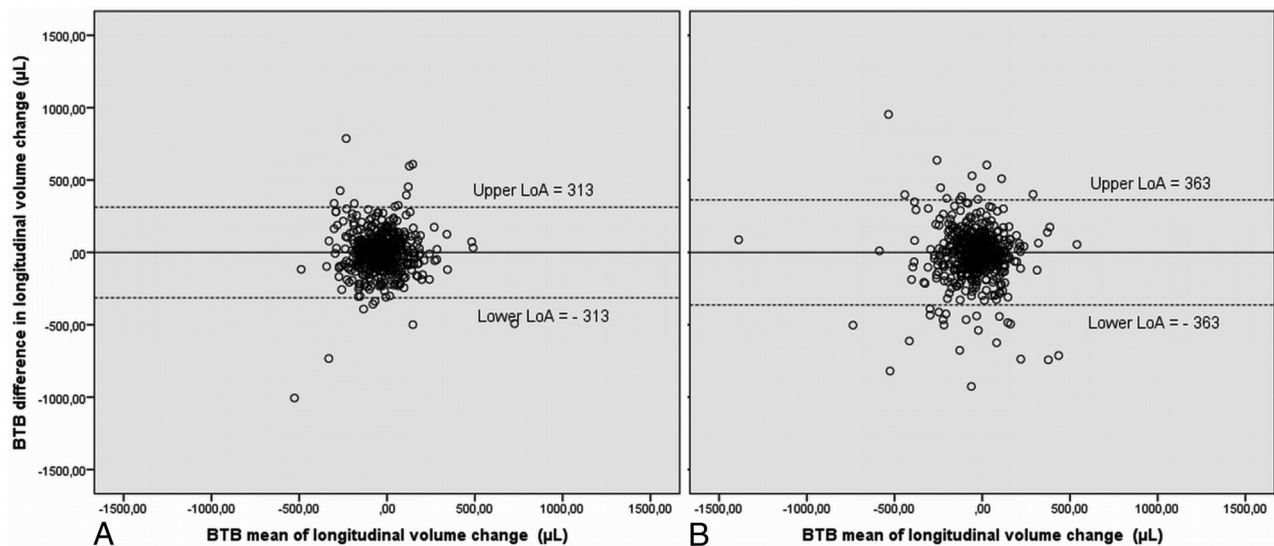


FIG 3. Bland-Altman plots for the left caudate nucleus, presented for the outcome measures of BTB difference in longitudinal volume change, to illustrate agreement for A, FreeSurfer and B, FIRST. Plots show the difference between the 2 measurements (ie, the “BTB difference”) along the vertical axis versus the mean of the 2 measurements along the horizontal axis. LoAs for FreeSurfer are obviously smaller (ie, better reproducibility) than those of FIRST.

Table 2: Median absolute BTB difference in longitudinal volume change for each deep GM structure for both hemispheres. Effect size, corresponding 95% confidence CI, and P values based on linear multilevel modelling are also presented

GM Structure	Software	Hemisphere	Median Absolute Difference, μl	90th Percentile	P Value	Effect	95% CI Lower	95% CI Upper
Caudate nucleus	FreeSurfer	Left	71.80	219.88	.003	0.82	0.72	0.93
	FIRST	Left	85.84	442.31				
	FreeSurfer	Right	74.20	253.22	.09	0.90	0.79	1.02
	FIRST	Right	87.26	264.69				
Putamen	FreeSurfer	Left	117.00	315.58	<.001	0.73	0.64	0.84
	FIRST	Left	156.39	412.49				
	FreeSurfer	Right	114.30	326.22	.003	0.77	0.64	0.91
Amygdala	FIRST	Right	142.44	406.71				
	FreeSurfer	Left	63.50	173.60	<.001	0.76	0.66	0.87
	FIRST	Left	84.38	260.14				
Globus pallidus	FreeSurfer	Right	77.00	196.54	<.001	0.73	0.64	0.83
	FIRST	Right	101.25	331.70				
	FreeSurfer	Left	59.60	180.56	.75	1.03	0.87	1.22
Thalamus	FIRST	Left	62.05	185.44				
	FreeSurfer	Right	60.00	164.38	.48	0.96	0.85	1.08
	FIRST	Right	60.06	200.23				
Thalamus	FreeSurfer	Left	100.90	315.08	.10	1.13	0.98	1.30
	FIRST	Left	91.76	262.30				
	FreeSurfer	Right	88.20	243.46	<.001	0.78	0.69	0.89
	FIRST	Right	114.81	308.71				

deep GM structures tended to have smaller BTB differences in longitudinal volume change and larger BTB differences in percentage volume change.

Effect Sizes

Effect sizes, based on the difference in the regression coefficient, corresponding P values, and 95% confidence intervals of comparison between segmentation by using FreeSurfer and FIRST are presented in Tables 2 and 3. The effect size in these Tables can be interpreted as the mean improvement of reproducibility in both longitudinal outcome variables when switching from FSL to FreeSurfer. For the outcome measure of the absolute BTB difference in longitudinal volume change, FreeSurfer performed significantly better than FIRST for the left and right amygdala (both $P < .001$),

left caudate nucleus ($P = .003$), left ($P < .001$) and right ($P = .003$) putamen, and right thalamus ($P < .001$). Concerning the outcome measure of the absolute BTB difference in longitudinal percentage volume change, FreeSurfer performed significantly better than FIRST for the left amygdala ($P = .02$), left ($P = .002$) and right ($P = .004$) caudate nucleus, and left putamen ($P < .001$). For the right amygdala and putamen, results are not presented because of lack of validity caused by failures in model fit.

Outliers

For the right amygdala, number of outliers were significantly different in all groups, when comparing 2 segmentation software packages ($P < .002$). This difference was not significant for other structures. Table 4 shows the number of excluded cases (extreme

Table 3: Median absolute BTB difference in longitudinal percentage volume change for each deep GM structure for both hemispheres. Effect size, corresponding 95% CI and P values based on linear multilevel modelling are also presented

GM Structure	Software	Hemisphere	Median Absolute Difference, %	90th Percentile	P Value	Effect	95% CI Lower	95% CI Upper
Caudate nucleus	FreeSurfer	Left	2.04	6.04	.002	0.09	0.02	0.41
	FIRST	Left	2.71	13.37				
	FreeSurfer	Right	2.08	6.92	.004	0.16	0.05	0.55
	FIRST	Right	2.63	8.15				
Putamen ^a	FreeSurfer	Left	2.48	6.87	<.001	0.07	0.02	0.29
	FIRST	Left	3.43	9.51				
	FreeSurfer	Right	2.46	6.77	NA	NA	NA	NA
	FIRST	Right	3.32	9.48				
Amygdala ^a	FreeSurfer	Left	6.14	17.64	.02	0.07	0.01	0.69
	FIRST	Left	6.60	21.29				
	FreeSurfer	Right	6.24	16.52	NA	NA	NA	NA
	FIRST	Right	8.63	28.95				
Globus pallidus	FreeSurfer	Left	4.32	13.04	.49	1.63	0.40	6.63
	FIRST	Left	3.65	10.92				
	FreeSurfer	Right	4.30	12.22	.94	0.95	0.29	3.17
	FIRST	Right	3.54	11.75				
Thalamus	FreeSurfer	Left	1.52	4.61	.33	0.62	0.23	1.64
	FIRST	Left	1.26	3.53				
	FreeSurfer	Right	1.41	3.82	.05	0.33	0.11	1.01
	FIRST	Right	1.59	4.31				

Note:—NA indicates not available.

^a For the right amygdala and putamen, results are not presented because of lack of validity caused by failures in the model fit.

Table 4: Number and proportion of excluded cases (extreme outliers) presented for each software for each deep GM structure

GM Structure	Software	Number of Outliers	% Within Segmentation	% Within Total Sample
Caudate nucleus	FreeSurfer	1	2.8	0.02
	FIRST	35	97.2	0.77
Putamen	FreeSurfer	0	0.0	0.00
	FIRST	37	100.0	0.82
Amygdala	FreeSurfer	33	17.8	0.73
	FIRST	152	82.2	3.38
Globus pallidus	FreeSurfer	21	39.6	0.46
	FIRST	32	60.4	0.71
Thalamus	FreeSurfer	0	0.0	0.00
	FIRST	29	100.0	0.64

outliers) for each deep GM structure and their proportion within each segmentation software used and the total sample size. The proportion of excluded cases was relatively small in the total sample of data; however it turned out to be more frequent when using FIRST compared with FreeSurfer.

Limits of Agreement

Based on our third method to evaluate reproducibility, values for the LoAs of FreeSurfer and FIRST derived from linear multilevel modeling are reported in Table 5. This analysis showed a visible trend for a better performance of FreeSurfer for both the measurement of longitudinal volume change and longitudinal percentage volume change, except for the globus pallidus, for which FIRST performed better. There was also a trend for an influence of the typical cross-sectional volume of a structure. Smaller deep GM structures showed smaller LoAs for longitudinal volume change measurement and larger LoAs for longitudinal percentage volume change.

DISCUSSION

Brain atrophy reflecting neurodegeneration and neuroaxonal damage is known to be an important characteristic of diseases like

AD and MS. In the current study, we investigated the reproducibility of volume change and percentage volume change measurement of 5 deep GM structures in a large multicenter dataset. To this end, we used 2 frequently used segmentation software packages, FreeSurfer and FIRST.

It is worth mentioning that FreeSurfer does provide a longitudinal pipeline to analyze multiple time points whereas FIRST only offers a cross-sectional analysis. Strikingly, for both software packages, the reproducibility error was comparable with the measured atrophy rates. Our results showed that BTB differences in 1-year percentage volume change (ranging from 1.26% for left thalamus to 8.63% for right amygdala) were roughly 1.5–3.5 times larger than the average atrophy rates of these deep GM structures (approximately 0.9% and 2.5%, respectively).

We used 3 different statistical methods that complement each other. Although reporting median and 90th percentile absolute differences alone is an easy and robust way to interpret results, statistical comparison in outcome measurements between methods of segmentation is not possible. Therefore, we next performed additional analytical statistics based on the difference in the regression coefficient. Finally, we used the method of determination of SEM, which provides a very precise way to map reproducibility and allows modeling of different sources of variability. This method is also proposed to be applied in determining agreement to map measurement error, an important measurement property in medicine.^{33,34,38} The sensitivity of this method is mainly attributable to the determination of specific variance components of a model, from which LoAs can be determined. In addition, the SEM method is a more suitable way for determining specific random variance in an outcome measure, which could provide additional information of the estimation of variance in a population. Using a large ADNI dataset makes such estimations more accurate. Another advantage of this method is that it is based on spread, contrary to the second regression-based method, and instead of

Table 5: SEMs and LoAs derived from variance component analysis out of a linear multilevel model for each deep GM structure

GM Structure	Software	SEM Longitudinal		SEM % Longitudinal	
		Volume Change, μl	LoA, μl	Volume Change	LoA, %
Caudate nucleus	FreeSurfer	112.89	312.92	2.90	8.05
	FIRST	125.83	348.79	3.76	10.42
Putamen	FreeSurfer	150.64	417.54	2.98	8.27
	FIRST	181.05	501.85	3.86	10.69
Amygdala	FreeSurfer	78.14	216.59	7.05	19.54
	FIRST	146.23	405.33	7.86	21.77
Globus pallidus	FreeSurfer	80.33	222.67	5.30	14.69
	FIRST	74.22	205.74	4.64	12.87
Thalamus	FreeSurfer	128.02	354.86	1.87	5.17
	FIRST	142.53	395.07	1.96	5.43

signed or absolute BTB differences, the clinical variables of interest (eg, volume change, percentage volume change) are modeled directly. This method for determining LoAs, however, is strongly affected by large outliers, and its procedure is much more costly and time-consuming.

Both methods of analytical statistics, namely determination of SEM with derived LoAs and the method based on difference in the regression coefficient, were carried out by using linear multilevel modeling. The general advantages of linear multilevel analysis are its flexibility in dealing with missing data, the ability to objectively include factors and covariates into 1 whole model, and a necessary applied correction for the dependency of data for measurements within the same subjects.^{39,40}

For both software packages, the reproducibility error was substantial compared with the measured atrophy (see Table 1 for the measured atrophy). However, FreeSurfer had better reproducibility compared with FIRST within the whole longitudinal outcome spectrum (except for globus pallidus), though the differences were not very large. The reproducibility was dependent on the structure baseline volume and also on the desired outcome measure (ie, volume change or percentage volume change). For example, compared with larger structures, smaller GM structures had smaller reproducibility errors for volume change and larger reproducibility errors for percentage volume change. For the structures measured in our study, when measuring the longitudinal volume change, the larger GM structures (putamen and thalamus) had BTB differences roughly twice as large as smaller structures (amygdala, globus pallidus), whereas for the outcome of longitudinal percentage volume change, this was reversed: here, larger structures outperformed smaller structures by approximately a factor of 5. A study on cross-sectional volume measurement by using FreeSurfer,⁴¹ reported generally larger relative scan-rescan errors for smaller structures. Such variability could cause poorer reproducibility of longitudinal volume change for smaller structures.

This poor reproducibility could be linked to the poor delineation of such brain structures by using automated software. To improve this, increase in the SNR and contrast-to-noise ratio (eg, by increasing the field strength or by further optimization of the acquisition) are recommended. In addition, multimodal segmentation, which includes other tissue information such as diffusion and susceptibility, could increase the accuracy and reproducibility of the segmentation and volume estimation.

Our study had some limitations. Because of the very large

number of segmentations performed, visual inspection of segmentation results was impractical. However, we used an automated method to exclude gross segmentation errors by using the BTB information. The very few occurring implausible outliers in our outcome measures were assumed to be caused by incorrect segmentations of 1 or more scans of that subject. To identify such gross outliers without excluding true atrophies, we applied a very wide cutoff criterion of 25% in longitudinal volume change or in percentage volume change compared with the baseline. As expected, the LoAs were very large when including the improbable outliers.

CONCLUSIONS

We provided quantitative information for 5 deep GM structures by using the widely used segmentation algorithms FreeSurfer and FIRST by 3 different methods of analysis. In general, FreeSurfer performance was better than that of FIRST. However, our results showed that BTB differences in 1-year percentage volume change were roughly 1.5–3.5 times larger than the atrophy rates of those deep GM structures. This suggests that longitudinal deep GM atrophy measures should be interpreted with caution. Finally, to provide a reliable additional biomarker, deep GM atrophy measurement techniques require substantially improved reproducibility, specifically when aiming for personalized medicine.

Disclosures: Marianne Jonker—UNRELATED: Consulting Fee or Honorarium: Vrije University Medical Center, Department of Epidemiology and Biostatistics, Comments: as a common practice, our department is paid for doing consultancy*. Ronald van Schijndel—OTHER RELATIONSHIPS: Image Analysis Center (IAC), Comments: partly working for the IAC, which is a contract research organization of the Vrije University Medical Center. Keith Cover—UNRELATED: Grants/Grants Pending: European Community FP7 Project, Comments: funded by the Neugrid4you project (grant agreement 2835262) from 2011 to 2015; studied methods for measuring atrophy of the brain in MRI*. Hugo Vrenken—UNRELATED: Grants/Grants Pending: Novartis Pharma, Comments: brain atrophy in MS; Teva Europe, Comments: brain atrophy in MS; Merck Serono, Comments: brain atrophy and lesions in MS*. *Money paid to the institution.

REFERENCES

1. Burns JM, Morris JC. **Neuropathology of Alzheimer's disease, non-demented aging and MCI.** In: *Mild Cognitive Impairment and Early Alzheimer's Disease: Detection and Diagnosis.* Hoboken: John Wiley & Sons; 2008:17–34
2. Iqbal K, Liu F, Gong CX, et al. **Mechanisms of tau-induced neurodegeneration.** *Acta Neuropathol* 2009;118:53–69 CrossRef Medline
3. Overk CR, Masliah E. **Pathogenesis of synaptic degeneration in Alzheimer's disease and Lewy body disease.** *Biochem Pharmacol* 2014; 88:508–16 CrossRef Medline

4. Stelmashook EV, Isaev NK, Genrikhs EE, et al. **Role of zinc and copper ions in the pathogenetic mechanisms of Alzheimer's and Parkinson's diseases.** *Biochemistry (Mosc)* 2014;79:391–96 CrossRef Medline
5. Mufson EJ, Binder L, Counts SE, et al. **Mild cognitive impairment: pathology and mechanisms.** *Acta Neuropathol* 2012;123:13–30 CrossRef Medline
6. Stephan BC, Hunter S, Harris D, et al. **The neuropathological profile of mild cognitive impairment (MCI): a systematic review.** *Mol Psychiatry* 2012;17:1056–76 CrossRef Medline
7. Maghzi AH, Revirajan N, Julian LJ, et al. **Magnetic resonance imaging correlates of clinical outcomes in early multiple sclerosis.** *Mult Scler Relat Disord* 2014;3:720–27 CrossRef Medline
8. Rudick RA, Fisher E, Lee JC, et al. **Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. Multiple Sclerosis Collaborative Research Group.** *Neurology* 1999;53:1698–704 CrossRef Medline
9. Sluiter JD, van der Flier WM, Karas GB, et al. **Whole-brain atrophy rate and cognitive decline: longitudinal MR study of memory clinic patients.** *Radiology* 2008;248:590–98 CrossRef Medline
10. Vercellino M, Masera S, Lorenzatti M, et al. **Demyelination, inflammation, and neurodegeneration in multiple sclerosis deep gray matter.** *J Neuropathol Exp Neurol* 2009;68:489–502 CrossRef Medline
11. Jiji S, Smitha KA, Gupta AK, et al. **Segmentation and volumetric analysis of the caudate nucleus in Alzheimer's disease.** *Eur J Radiol* 2013;82:1525–30 CrossRef Medline
12. Leung KK, Bartlett JW, Barnes J, et al. **Cerebral atrophy in mild cognitive impairment and Alzheimer disease: rates and acceleration.** *Neurology* 2013;80:648–54 CrossRef Medline
13. Macfarlane MD, Looi JC, Walterfang M, et al. **Executive dysfunction correlates with caudate nucleus atrophy in patients with white matter changes on MRI: a subset of LADIS.** *Psychiatry Res* 2013;214:16–23 CrossRef Medline
14. Miller MI, Younes L, Ratnanather JT, et al. **Amygdalar atrophy in symptomatic Alzheimer's disease based on diffeomorphic registration: the BIOCARD cohort.** *Neurobiol Aging* 2015;36 Suppl 1:S3–S10 CrossRef Medline
15. Skup M, Zhu H, Wang Y, et al. **Sex differences in grey matter atrophy patterns among AD and aMCI patients: results from ADNI.** *Neuroimage* 2011;56:890–906 CrossRef Medline
16. Štěpán-Buksakowska I, Szabó N, Hořínek D, et al. **Cortical and subcortical atrophy in Alzheimer disease: parallel atrophy of thalamus and hippocampus.** *Alzheimer Dis Assoc Disord* 2014;28:65–72 CrossRef Medline
17. Yi HA, Möller C, Dieleman N, et al. **Relation between subcortical grey matter atrophy and conversion from mild cognitive impairment to Alzheimer's disease.** *J Neurol Neurosurg Psychiatry* 2016;87:425–32 CrossRef Medline
18. Batista S, Zivadinov R, Hoogs M, et al. **Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis.** *J Neurol* 2012;259:139–46 CrossRef Medline
19. Ramasamy DP, Benedict RH, Cox JL, et al. **Extent of cerebellum, subcortical and cortical atrophy in patients with MS: a case-control study.** *J Neurol Sci* 2009;282:47–54 CrossRef Medline
20. Shiee N, Bazin PL, Zackowski KM, et al. **Revisiting brain atrophy and its relationship to disability in multiple sclerosis.** *PLoS One* 2012;7:e37049 CrossRef Medline
21. Hanganu A, Bedetti C, Degroot C, et al. **Mild cognitive impairment is linked with faster rate of cortical thinning in patients with Parkinson's disease longitudinally.** *Brain* 2014;137:1120–29 CrossRef Medline
22. Mak E, Bergsland N, Dwyer MG, et al. **Subcortical atrophy is associated with cognitive impairment in mild Parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis.** *AJNR Am J Neuroradiol* 2014;35:2257–64 CrossRef Medline
23. Nemmi F, Sabatini U, Rascol O, et al. **Parkinson's disease and local atrophy in subcortical nuclei: insight from shape analysis.** *Neurobiol Aging* 2015;36:424–33 CrossRef Medline
24. Kishi T, Matsunaga S, Oya K, et al. **Protection against brain atrophy by anti-dementia medication in mild cognitive impairment and Alzheimer's disease: meta-analysis of longitudinal randomized placebo-controlled trials.** *Int J Neuropsychopharmacol* 2015;18:pyv070 CrossRef Medline
25. Patenaude B, Smith SM, Kennedy DN, et al. **A Bayesian model of shape and appearance for subcortical brain segmentation.** *Neuroimage* 2011;56:907–22 CrossRef Medline
26. Fischl B, Salat DH, Busa E, et al. **Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain.** *Neuron* 2002;33:341–55 CrossRef Medline
27. Fischl B, van der Kouwe A, Destrieux C, et al. **Automatically parcellating the human cerebral cortex.** *Cereb Cortex* 2004;14:11–22 CrossRef Medline
28. Reuter M, Rosas HD, Fischl B. **Highly accurate inverse consistent registration: a robust approach.** *Neuroimage* 2010;53:1181–96 CrossRef Medline
29. Reuter M, Schmansky NJ, Rosas HD, et al. **Within-subject template estimation for unbiased longitudinal image analysis.** *Neuroimage* 2012;61:1402–18 CrossRef Medline
30. Mueller SG, Weiner MW, Thal LJ, et al. **Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI).** *Alzheimers Dement* 2005;1:55–66 CrossRef Medline
31. Cover KS, van Schijndel RA, van Dijk BW, et al. **Assessing the reproducibility of the SienaX and Siena brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans.** *Psychiatry Res* 2011;193:182–90 CrossRef Medline
32. Jack CR Jr., Bernstein MA, Fox NC, et al. **The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods.** *J Magn Reson Imaging* 2008;27:685–91 CrossRef Medline
33. Carstensen B, Simpson J, Gurrin LC. **Statistical models for assessing agreement in method comparison studies with replicate measurements.** *Int J Biostat* 2008;4:Article 16 CrossRef
34. de Vet HCW, Terwee CB, Mokkink LB, et al. *Measurement in Medicine: A Practical Guide.* New York: Cambridge University Press; 2011
35. Mulder ER, de Jong RA, Knol DL, et al. **Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST.** *Neuroimage* 2014;92:169–81 CrossRef Medline
36. Bland JM, Altman DG. **Statistical methods for assessing agreement between two methods of clinical measurement.** *Lancet* 1986;1:307–10 CrossRef Medline
37. Euser AM, Dekker FW, le Cessie S. **A practical approach to Bland-Altman plots and variation coefficients for log transformed variables.** *J Clin Epidemiol* 2008;61:978–82 CrossRef Medline
38. Terwee CB, Bot SD, de Boer MR, et al. **Quality criteria were proposed for measurement properties of health status questionnaires.** *J Clin Epidemiol* 2007;60:34–42 CrossRef Medline
39. Twisk JWR. *Applied Multilevel Analysis: A Practical Guide for Medical Researchers.* New York: Cambridge University Press; 2006
40. Twisk JWR. *Applied Longitudinal Data Analysis for Epidemiology.* New York: Cambridge University Press; 2013
41. Jovicich J, Czanner S, Han X, et al. **MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths.** *Neuroimage* 2009;46:177–92 CrossRef Medline