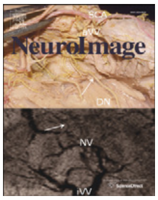
NeuroImage 92 (2014) 169–181

Contents lists available at ScienceDirect

NeuroImage

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



Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST

Emma R. Mulder [a,b], Remko A. de Jong [a,b], Dirk L. Knol [c], Ronald A. van Schijndel [a,d], Keith S. Cover [e], Pieter J. Visser [f], Frederik Barkhof [a,b], Hugo Vrenken [b,e,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] *Image Analysis Center, VU University Medical Center, Amsterdam, The Netherlands*
[b] *Department of Radiology, VU University Medical Center, Amsterdam, The Netherlands*
[c] *Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands*
[d] *Department of Information and Communication Technology, VU University Medical Center, Amsterdam, The Netherlands*
[e] *Department of Physics and Medical Technology, VU University Medical Center, Amsterdam, The Netherlands*
[f] *Department of Neurology, VU University Medical Center, Amsterdam, The Netherlands*

ARTICLE INFO

Article history:
Accepted 31 January 2014
Available online 9 February 2014

Keywords:
Hippocampus
Manual segmentation
Automatic segmentation
Magnetic resonance imaging
Mild cognitive impairment
Alzheimer's disease

ABSTRACT

Background: To measure hippocampal volume change in Alzheimer's disease (AD) or mild cognitive impairment (MCI), expert manual delineation is often used because of its supposed accuracy. It has been suggested that expert outlining yields poorer reproducibility as compared to automated methods, but this has not been investigated.

Aim: To determine the reproducibilities of expert manual outlining and two common automated methods for measuring hippocampal atrophy rates in healthy aging, MCI and AD.

Methods: From the Alzheimer's Disease Neuroimaging Initiative (ADNI), 80 subjects were selected: 20 patients with AD, 40 patients with mild cognitive impairment (MCI) and 20 healthy controls (HCs). Left and right hippocampal volume change between baseline and month-12 visit was assessed by using expert manual delineation, and by the automated software packages FreeSurfer (longitudinal processing stream) and FIRST. To assess reproducibility of the measured hippocampal volume change, both back-to-back (BTB) MPRAGE scans available for each visit were analyzed. Hippocampal volume change was expressed in μL, and as a percentage of baseline volume. Reproducibility of the 1-year hippocampal volume change was estimated from the BTB measurements by using linear mixed model to calculate the limits of agreement (LoA) of each method, reflecting its measurement uncertainty. Using the delta method, approximate p-values were calculated for the pairwise comparisons between methods. Statistical analyses were performed both with inclusion and exclusion of visibly incorrect segmentations.

Results: Visibly incorrect automated segmentation in either one or both scans of a longitudinal scan pair occurred in 7.5% of the hippocampi for FreeSurfer and in 6.9% of the hippocampi for FIRST. After excluding these failed cases, reproducibility analysis for 1-year percentage volume change yielded LoA of ±7.2% for FreeSurfer, ±9.7% for expert manual delineation, and ±10.0% for FIRST. Methods ranked the same for reproducibility of 1-year μL volume change, with LoA of ±218 μL for FreeSurfer, ±319 μL for expert manual delineation, and ±333 μL for FIRST. Approximate p-values indicated that reproducibility was better for FreeSurfer than for manual or FIRST, and that manual and FIRST did not differ. Inclusion of failed automated segmentations led to worsening of reproducibility of both automated methods for 1-year raw and percentage volume change.

Conclusion: Quantitative reproducibility values of 1-year microliter and percentage hippocampal volume change were roughly similar between expert manual outlining, FIRST and FreeSurfer, but FreeSurfer reproducibility was statistically significantly superior to both manual outlining and FIRST after exclusion of failed segmentations.

[*] Corresponding author at: VU University Medical Center, Department of Radiology and Department of Physics and Medical Technology, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands.
E-mail address: H.Vrenken@vumc.nl (H. Vrenken).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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 is available at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1053-8119/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2014.01.058

Introduction

In amnestic mild cognitive impairment (MCI), a prodromal stage of Alzheimer's disease (AD), hippocampal volume loss has been found to exceed that in age-matched controls (Drago et al., 2011), and to be related to risk of subsequent conversion to AD (Drago et al., 2011). Therefore, hippocampal volume change measurement from magnetic resonance (MR) images has been a major focus of recent studies, and a common outcome measure in clinical trials on AD and MCI (Ard and Edland, 2011; Schott et al., 2010).

To date, manual outlining of the hippocampus by experts in neuroanatomy has been the standard approach (Barnes et al., 2008; Boccardi et al., 2011). The strength of expert manual delineation is its supposed accuracy in correctly identifying the hippocampi (Boccardi et al., 2011), which is guaranteed by having a trained expert outline the hippocampi at each slice. However, the trained experts introduce both inter-rater and intra-rater variability of the manual outlines (Boccardi et al., 2011), which are absent for automated methods without user intervention. Therefore, it has been suggested that manual methods have poorer reproducibility than automated methods (Dewey et al., 2010; Doring et al., 2011; Duchesne et al., 2002; Kennedy et al., 2009). Surprisingly, this hypothesis has so far not been tested directly.

Therefore, in this study, we directly compared the reproducibility of 1-year hippocampal volume change measurement between expert manual delineation, and two frequently used fully automated methods: FreeSurfer (Reuter et al., 2012), and FIRST (Patenaude et al., 2011). The study was restricted to these three methods in order to capture the extremes of possible approaches to hippocampal segmentation: fully manual on the one hand, and fully automated (FreeSurfer and FIRST) on the other. To assess reproducibility, we employed the unique set of “back-to-back” (BTB) MPRAGE scans that are available as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Jack et al., 2008; Mueller et al., 2005; Weiner et al., 2012). These within-session scan–rescan scan pairs, acquired with only a few seconds to minutes in between scans, are very similar to each other but not exactly identical. Therefore, they are ideal for performing a reproducibility analysis of automated methods, as was previously done for whole-brain atrophy by Cover et al. (2011). We calculated the median absolute difference, as well as the limits of agreement obtained from linear mixed model analysis.

Materials and methods

ADNI

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu) (Jack et al., 2008; Mueller et al., 2005; Weiner et al., 2012). The ADNI study was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. 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 assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

Subjects

From the ADNI database, baseline and month-12 visits of 80 subjects were selected, of which 20 with diagnosed AD, 20 healthy controls and 40 MCI patients (with baseline visits between September 2005 and August 2007). Two subgroups of MCI patients were a priori selected based on CSF profile, using the AD-positive cut-off value of tau/Aβ_{1-42} ≥ 0.39 determined by Shaw et al., 2009). We selected 20 MCI patients with an AD-positive CSF profile (MCI-P; tau/A$\beta_{1-42} \geq 0.39$), and 20 MCI patients with an AD-negative profile (MCI-N; tau/A$\beta_{1-42} < 0.39$). Subjects in the healthy control group all had a tau/Aβ_{1-42} value of <0.39. In the AD group, all but one subject had a tau/Aβ_{1-42} value of >0.39. Based on the updated clinical database of January 2013, two healthy controls converted to MCI, and two MCI-N patients and eight MCI-P patients converted to AD. Table 1 lists the subject characteristics.

MRI acquisition

3D T1 weighted MPRAGE scans were acquired at 1.5 T at multiple sites using scanners from various vendors (Philips, Siemens and GE). For both the baseline visit and month-12 visit, two raw 3DT1 DICOM image series per visit were downloaded from the public ADNI database (http://www.loni.ucla.edu/ADNI/Data/index.shtml). Visual inspection was performed to ensure that for each visit of the selected subjects, both the “back-to-back” 3DT1 image volumes were of good quality. Images were not post-processed beyond scanner default corrections. More detailed information on MR acquisition in ADNI has been published previously (Jack et al., 2008).

Hippocampal volumetry

Manual segmentation

For manual segmentation of the hippocampus we followed the standard operating procedure that is used in clinical trials at the Image Analyses Center (IAC, Amsterdam). Baseline scans were reformatted in a plane perpendicular to the long axis of the left

Table 1
Subject demographics.

	CTR (n = 20)	MCI-N (n = 20)	MCI-P (n = 20)	AD (n = 20)
Age, years	75.7 (6.1)	74.3 (7.3)	73 (6.7)	72.6 (6.9)
% male	60%	65%	65%	55%
MMSE score at screening	29.2 (1.14)	27.1 (1.74)	27.7 (1.84)	24.4 (1.18)
t-Tau/Aβ[a] median [range]	0.23 [0.13; 0.31]	0.25 [0.14; 0.34]	0.87 [0.47; 1.46]	1.04 [0.23; 2.03]
Conversions[b]	2 → MCI	2 → AD	8 → AD	n.a.
Baseline volume FIRST	3568 (459)	3433 (573)	3258 (588)	2972 (566)
Baseline volume manual	3469 (395)	3255 (551)	3089 (491)	2893 (570)
Baseline volume FreeSurfer	3558 (408)	3276 (564)	3069 (535)	2826 (576)

Data are presented as mean (SD) unless indicated otherwise.
HC: healthy control; MCI-N: mild cognitive impaired with AD-negative t-tau/Aβ profile; MCI-P: mild cognitive impaired with AD-positive t-tau/Aβ profile; AD: Alzheimer's disease; MMSE: Mini-Mental State Examination.
[a] Adapted from Shaw et al. (2009), with ≥.39 for AD positive.
[b] Until January 2013.

hippocampus, using 2 mm thick slices with the original in-plane resolution and sinc interpolation. Month-12 scans were rigid body registered to the reformatted baseline scan, using the same resolution and sinc interpolation. Both right and left hippocampi were manually segmented by one trained technician (F.C. van D.) at the Image Analysis Center, using in-house-developed software (Show_Images v3.7.1.0, VU University Medical Center) on a Linux Ubuntu workstation. The hippocampus was segmented according to the criteria described by van de Pol et al. (2007). Manually segmenting both hippocampi on one scan took approximately three hours. Hippocampal volume (in μL) was calculated by multiplying the total area of all ROIs of each hippocampus by slice thickness. The technician was blinded to the diagnosis, but not to the visit, because the workflow demands the follow-up visit to be compared with the previous visit. However, the first and second scan pairs of the BTB pairs were assigned in a random order to avoid any training effect.

Automated methods

We included two frequently used fully automated methods for hippocampal volume change measurement: FreeSurfer (Fischl et al., 2002, 2004; Reuter et al., 2012), a software package for subcortical segmentation and cortical parcellation, and FIRST (FMRIB's Integrated Registration and Segmentation Tool) (Patenaude et al., 2011), an automated tool for subcortical segmentation which is part of the FMRIB Software Library (FSL). Both are freely available for academic use. For the automated methods, DICOM images provided by ADNI were converted to NIfTI-format. All automated hippocampal segmentations were performed on a 64-bit Linux machine.

FreeSurfer

Automated hippocampal segmentation was done with FreeSurfer 5.1.0 (http://surfer.nmr.mgh.harvard.edu/) using the longitudinal processing stream, which requires cross-sectional processing of each visit, followed by joint longitudinal processing of the longitudinal scan pair. Each cross-sectional job took approximately 25 h using the default processing stream (recon-all -all). Constructing templates from both time points for each subject and processing them longitudinally using the longitudinal processing stream (recon-all base and recon-all long respectively) additionally required approximately 40 h for each longitudinal scan pair.

The stages of FreeSurfer's volume-based subcortical stream are fully described elsewhere (Fischl et al., 2002, 2004). Briefly, an affine registration with Talairach space is followed by an initial volumetric labeling and correction for variation in intensity due to the B1 bias field. After this, a high dimensional nonlinear volumetric alignment to the Talairach atlas is performed, followed by pre-processing, and finally the volume is labeled. Hippocampal volume was calculated by multiplying the number of voxels by the voxel volume. In the longitudinal processing stream, instead of registration to Talairach space, an unbiased within-subject template space and average image (Reuter et al., 2012) is created using robust, inverse consistent registration (Reuter et al., 2010). No manual editing was performed at any stage of the FreeSurfer processing stream.

FSL-FIRST

Automated hippocampal segmentation was done with FSL-FIRST v.4.1.5 (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST) using the default settings. The stages of subcortical segmentation by FIRST have been described in detail by Patenaude et al. (2011). In short, a two-stage affine registration to MNI152 space is used to obtain normalized spatial coordinates for the vertices of the meshes used by FIRST to model the outer surface of the brain structures. FIRST segmentation is based on shape and appearance models that have been constructed from manually segmented images. Appearance refers to normalized intensities along the surface normals, which are sampled and modeled. Shape refers to the normalized coordinates of the vertices composing the mesh, which are expressed as a mean with modes of variation (principal components). Based on the learned models, FIRST searches through linear combinations of shape modes of variation for the most probable shape instance given the observed intensities.

For quantification of hippocampal volumes, hippocampus meshes were converted to binary voxel ROIs using FIRST tools, with a boundary correction using FAST (Zhang et al., 2001), and hippocampal volumes were then calculated by multiplying the number of voxels by the voxel volume. Importantly, FIRST allows volumetric analysis at the voxel level while taking neighboring structures into account. Therefore, in the current study we used the hippocampal volume obtained through the segmentation of all subcortical structures by applying the run-first-all script, which takes about 50 min per scan, and subsequently calculating the volumes of the boundary-corrected hippocampus ROIs.

Visual inspection of segmentations

All hippocampal segmentation results, for the automated methods as well as the manual method, were visually inspected (by author E.R.M.). Hippocampal segmentation was considered to have failed when a substantial part of the hippocampus was identified as amygdala, ventricle or cortex, or when a substantial part of the amygdala, ventricle or cortex was identified as hippocampus, on two or more slices. In order to obtain some insight in the possible causes of segmentation failures, we visually inspected these MPRAGE images for the degree of whole-brain atrophy, the degree of ventricle dilation, and the amount of (MPRAGE-visible) white matter abnormalities. For each of these aspects we also assessed whether they were clearly asymmetrical.

Statistical analysis

The flow chart in Fig. 1 illustrates how change measures were compared between BTB scan pairs. Two longitudinal scan pairs were composed for each subject, and hippocampal volume change was measured for each longitudinal scan pair using all three methods (FIRST, manual and FreeSurfer). The hippocampal volume change in μL was calculated as $V(M12) - V(M0)$, in which $V(M12)$ represents the month12 volume in μL, and $V(M0)$ represents the baseline volume in μL. The percentage hippocampal volume change was calculated as: $100 * [V(M12) - V(M0)] / V(M0)$.

Reproducibility of each method was quantified in two ways. Firstly, by simply calculating the median absolute difference between the results obtained from the two longitudinal scan pairs for each subject. This is a simple but commonly reported measure that allows comparison with other studies, but does not provide information about the tails of the distribution of the difference. Secondly, therefore, we calculated the limits of agreement (LoA) for each hippocampal volumetry method. To this end, linear mixed modeling was performed using SAS v9.2. Separate linear mixed model analyses were performed for the two outcome measures: hippocampal volume change in μL, and percentage hippocampal volume change. For each outcome measure, all results were analyzed in a single model to provide the most objective comparison between the three hippocampal volumetry methods. Factors in the linear mixed model analyses were hippocampal volumetry method (M), diagnostic group (G), persons (P), hemisphere (H), and scan pair (S). Fixed effects were G, H and M, and all interactions between these three factors. For each outcome variable, three versions of the statistical model, with increasing levels of complexity, were evaluated, and the model with the lowest value for the Akaike Information Criterion (AIC) was selected. In all analyses, the lowest value for AIC was observed for the model with the highest complexity, and therefore we only describe this latter model in detail here. This model consisted, in addition to the fixed effects listed above, of the following nested relations, all modeled as random effects: First, persons nested in diagnostic group, which, using conventional notation (Searle et al., 1992) and the factor symbols defined above, can be written as P:G. Using the same notation, the other effects included were PH:G, PM:G,

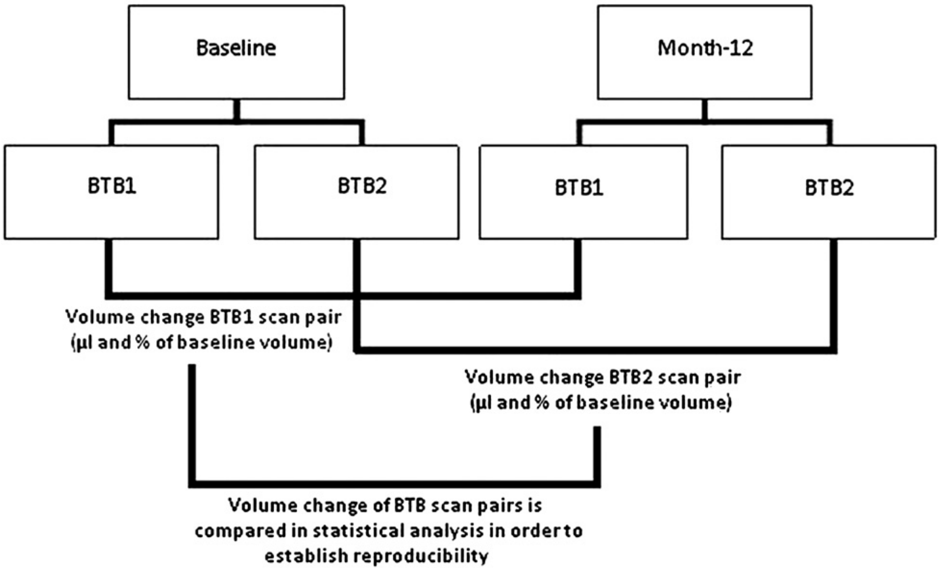

Baseline
Month-12
BTB1
BTB2
BTB1
BTB2
Volume change BTB1 scan pair
(μl and % of baseline volume)
Volume change BTB2 scan pair
(μl and % of baseline volume)
Volume change of BTB scan pairs is
compared in statistical analysis in order to
establish reproducibility

Fig. 1. Flowchart of hippocampal volume change analyses. At each timepoint (baseline or month-12), two back-to-back (BTB) MPRAGE scans were available. One of the baseline BTB scans was combined with one of the month-12 BTB scans to produce the first one-year hippocampal volume change analysis. The other baseline BTB scan was combined with the other month-12 BTB scan to produce the second one-year hippocampal volume change analysis. Both measurements were used to analyze reproducibility of the measured hippocampal volume change for each measurement method.

PHM:G, S:PG, SH:PG, and SM:PG. Finally, the residual effects (E) of the model can be interpreted as SHM:PG. All variance components containing S were allowed to vary with M.

From the linear mixed model, the interscan standard error of measurement (SEM) for each method was obtained from the variance components (σ^2) as the conditional variance of the change variable D, given P, G, H and M using Eq. (1):

$$\mathrm{SEM}^2 = \mathrm{Var}\left(\mathrm{D}_{\mathrm{pghms}}|\mathrm{p},\mathrm{g},\mathrm{h},\mathrm{m}\right) = \sigma^2_{\mathrm{S:PG}} + \sigma^2_{\mathrm{SH:PG}} + \sigma^2_{\mathrm{SM:PG}} + \sigma^2_{\mathrm{E}}. \tag{1}$$

The value of SEM calculated for each method was used to calculate the limits of agreement (LoA) using Eq. (2):

$$\mathrm{LoA} = \pm 1.96 \times \mathrm{SEM}\sqrt{2}. \tag{2}$$

In a typical clinical trial setting, segmentation results from automated methods would be visually inspected, and either rejected or edited if the segmentation result was incorrect. Therefore, we performed all primary linear mixed model analyses for the subset of measurements that remained after excluding failed segmentations. However, in very large datasets, it may be considered too costly to perform the visual inspections. Therefore, to quantify the effect of including the failed segmentations, we repeated all linear mixed model analyses using the total set of subjects, i.e. including the failed segmentations. We also repeated calculation of the median absolute differences for this case.

To obtain an indication of whether any observed differences between the LoAs of the three methods from our primary analysis exceeded chance level, we calculated approximate p-values for the pairwise comparisons of LoAs between methods, using the delta method (Hoef, 2012). Using Eq. (1), we calculated pairwise SEM^2 log differences between each pair of methods, and estimated the variance of each pairwise SEM^2 difference using the covariance matrix of the estimated variance components from the linear mixed model analysis. By dividing these two, a Z value was calculated for each pairwise SEM^2 difference, and under the assumption that this Z follows a normal distribution the p-value for the pairwise comparison was derived.

Finally, to assess the effect of the reproducibility on the relative sample size requirements for each hippocampal atrophy measurement method, we calculated the ratio of required sample sizes using the relative efficiency of each method (say, A) with respect to one of the two other methods (say, B). The relative efficiencies can be calculated from their method-specific inter-scan ICCs, which can be expressed in terms of their method-specific SEMs, as follows:

$$\frac{\mathrm{N_B}}{\mathrm{N_A}} = \mathrm{Efficiency}_{\mathrm{AvsB}} = \frac{\mathrm{ICC_A}}{\mathrm{ICC_B}} = \frac{\left(1-\frac{\mathrm{SEM}_\mathrm{A}^2}{\sigma^2_{\mathrm{total}}}\right)}{\left(1-\frac{\mathrm{SEM}_\mathrm{B}^2}{\sigma^2_{\mathrm{total}}}\right)}. \tag{3}$$

To illustrate agreement in volume change measurement between the two longitudinal scan pairs, Bland–Altman mean-difference plots (Bland and Altman, 1986) with previously described modifications (Euser et al., 2008) were generated in SPSS 16.0, both for volume change in μL and for volume change expressed as a percentage of the baseline volume. For description of the study sample, we also report the baseline volumes of left and right hippocampi obtained with each hippocampus volumetry method.

Results

Failed segmentations

Typical correct segmentation results for each of the three methods are illustrated in Fig. 2. Fig. 3 illustrates failed segmentations for both FreeSurfer and FIRST in the same subject. Fig. 4 shows examples of common small segmentation errors by both automated methods that were not considered to be severe enough to warrant classification of the segmentation as failed, but did occur in many cases. The two main errors of this kind were the inclusion of pockets of CSF within the hippocampus, illustrated in Fig. 4A for FIRST, and inclusion of cerebrospinal fluid (CSF) along the boundaries of the hippocampus, illustrated in Fig. 4B for FreeSurfer. In total, for cross-sectional analyses, 34 out of 640, i.e. 5.3%, of the hippocampal segmentations had failed for FIRST, and 41/640 (6.4%) of the hippocampal segmentations had failed for FreeSurfer. Failure of the longitudinal analyses, i.e. failed segmentation for at least one of the two timepoints, occurred in 22/320 (6.9%) of the hippocampal analyses for FIRST, and in 24/320 (7.5%) of the hippocampal analyses

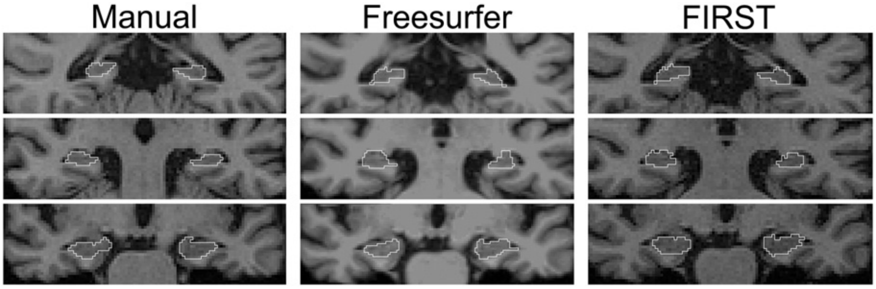


Fig. 2. Example of typical hippocampal segmentations, with posterior (top row), intermediate (middle row) and anterior (bottom row) sections through the long plane of the hippocampus in a female MCI patient (MCI-P group), age 68 years. Left column shows results for manual hippocampal segmentation, middle column for FreeSurfer, and right column for FIRST.

for FreeSurfer. For the manual measurement, no segmentations had failed.

We investigated possible causes for the failed hippocampus segmentations for FIRST (34/640) and FreeSurfer (41/640). To do so we rated whole-brain atrophy, ventricle dilation and white matter abnormalities visually. It should be noted that the total of 640 hippocampi were derived from just 80 subjects, with two scans at each of the two timepoints, and (obviously) two hippocampi. None of the visual ratings changed between the two timepoints for any subject. For FIRST, the 34 failed hippocampus segmentations were derived from 8 unique subjects, who had scans in which whole-brain atrophy, ventricle dilation and white matter abnormalities all ranged from absent to moderate to severe. A subset of just three subjects with >4 fails each, contributed the majority of fails (18 fails). In these three subjects also, whole-brain atrophy, ventricle dilation and white matter abnormalities all ranged from absent to moderate to severe, with one subject displaying asymmetry on all aspects, and one subject displaying moderately asymmetric atrophy.

A similar distribution was observed for FreeSurfer: the 41 failed hippocampus segmentations were derived from 13 unique subjects, who had scans in which whole-brain atrophy, ventricle dilation and white matter abnormalities all ranged from absent to moderate to severe. A subset of just four subjects with >4 fails each, contributed the majority of fails (27 fails). In these four subjects, whole-brain atrophy and ventricle dilation ranged from moderate to severe (moderately asymmetric in two subjects), while white matter abnormalities ranged from absent to moderate (no asymmetries).

Notably, of the 13 subjects with failed segmentations for FreeSurfer, only 3 subjects also had any failed segmentations for FIRST (with 2, 3 and 6 failures respectively); the total number of unique subjects with any failures for either FreeSurfer or FIRST or both was 18. Finally, the robustness of both methods is illustrated by the following: in this set of 18 subjects with at least one failed segmentation, there were 7 subjects with just 1 failed hippocampus segmentation for FreeSurfer and no failures for FIRST; in these subjects, all possible degrees of whole-brain atrophy, ventricle dilation and white matter abnormalities were also observed. As may be expected, segmentation failures were slightly more frequent in AD (7 out of 20 patients had failures) than in MCI (7 out of 40 patients had failures) or healthy controls (4 out of 20 subjects had failures).

Longitudinal hippocampal volume change results

Baseline hippocampal volume (Table 1) and hippocampal volume change over the 1-year interval (Table 2) were in the expected range for each clinical group. Table 2 provides, for each hippocampal volumetry method, the median and interquartile range of longitudinal hippocampal volume change for each subgroup. In view of limited journal space, we here present results averaged across both longitudinal scan pairs and across both hippocampi. More detailed information, with values for separate longitudinal scan pairs and for left and right hippocampi separately is provided in the Supplementary material. To illustrate our "raw" results in more detail, Fig. 5 shows percentage hippocampal volume change of the left hippocampus, for one of the longitudinal scan pairs, in the shape of a box-plot. All three methods show the expected sequence of HC–MCI–AD, with respect to percentage volume change, with FreeSurfer and FIRST showing higher variability in their whiskers among subject groups. Although obviously there was some variation between the two longitudinal scan pairs, as well as between the left and right hippocampi, the patterns observed for the other cases were highly similar to those shown in Fig. 5. In order to allow comparison between the three hippocampal volumetry methods, Fig. 5 and Tables 2 and 3 present data from all subjects, including the segmentations that were considered to have failed.

Reproducibility of longitudinal hippocampal volume change measurements

We analyzed reproducibility of the longitudinal volume change measurement for all three methods, first, by determining limits of agreement (LoA) from a joint linear mixed model analysis. FreeSurfer exhibited the smallest LoA for raw hippocampal volume change as well as for percentage volume change (±218 μL and ±7.2%, respectively), followed by the manual method (±319 μL and ±9.7%, respectively), and then by FIRST (±333 μL and ±10.0%, respectively). Note that these

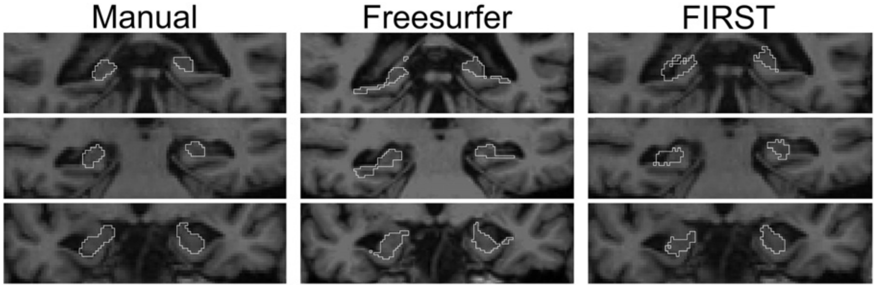


Fig. 3. Example of failed automated hippocampal segmentations, with posterior (top row), intermediate (middle row) and anterior (bottom row) sections through the long plane of the hippocampus in a male AD patient, age 74 years. Left column shows results for manual hippocampal segmentation, middle column for FreeSurfer, and right column for FIRST.

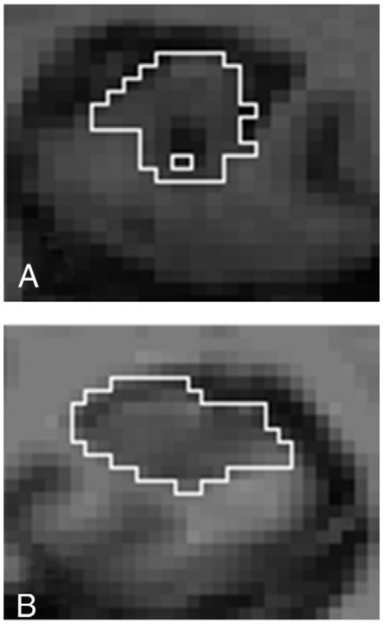


Fig. 4. Examples of consistent small errors in automated methods that were noted, but were considered not severe enough to warrant classification of the segmentation as failed. Panel (A) shows failure of FIRST to entirely exclude pockets of CSF from hippocampal segmentation. Panel (B) shows misidentification by FreeSurfer of CSF as hippocampus along the boundaries of the hippocampus.

results were obtained by treating failed segmentations as missing data in the linear mixed model. These results are summarized in Table 3 and further illustrated by the Bland–Altman mean-difference plots for raw hippocampal volume change (Fig. 6) and percentage volume change (Fig. 7). In both Figs. 6 and 7, panels show results for FIRST (A: left hippocampus, B: right hippocampus), manual segmentation (C: left hippocampus, D: right hippocampus), and FreeSurfer (E: left hippocampus, F: right hippocampus). Dashed lines indicate the LoA as derived from the linear mixed model when failed segmentations were treated as missing data. Second, we also determined median absolute differences directly from the data without modeling. In this case, to allow a fair comparison between the three methods, we excluded all hippocampi for which any of the segmentations had failed. The median absolute differences displayed similar trends to the LoA: median absolute differences were for FreeSurfer: left: 72.5 μL, 2.5%, right: 77.5 μL, 2.5%; for manual: left 92 μL, 2.5%, right: 105 μL, 3.6%; for FIRST: left: 89 μL, 2.8%, right: 121 μL, 3.6%.

When including the failed segmentations of the automated methods in the statistical model, reproducibility of hippocampal volume change measurement was substantially worse for the automated methods, with FreeSurfer having LoAs of ± 313 μL and ± 23.1%, and FIRST LoAs of ± 478 μL and ± 18.8%, although it should be noted that these numbers should be interpreted cautiously because of the large deviations of some of the outlier values. Due to the fact that LoAs were determined from joint linear mixed model analysis of all results, inclusion of the failed segmentations for the automated segmentations also induced a small change in the LoA values obtained for the manual method; these became ± 313 μL and ± 9.5%. As expected, the median absolute differences were less affected by including the failed segmentations, and these became for FreeSurfer: left: 73 μL, 2.6%, right: 84 μL, 2.6%; for manual: left 100 μL, 2.9%, right: 106 μL, 3.9%; and for FIRST: left: 85 μL, 2.6%, right: 121 μL, 3.6%.

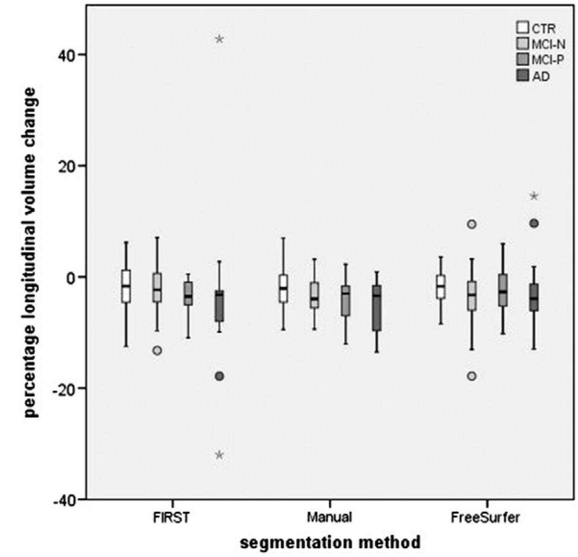


Fig. 5. Box-plot comparing percentage longitudinal volume change between the four different subgroups as measured by FIRST, manual method and FreeSurfer for the left hippocampus. CTR: healthy control; MCI-N: mild cognitive impairment with AD-negative t-tau/Aβ profile; MCI-P: mild cognitive impairment with AD-positive t-tau/Aβ profile; AD: Alzheimer's disease.

To obtain an indication of whether the observed differences exceeded chance level, we calculated approximate p-values for the pairwise comparisons of LoAs between methods, using the delta method (Hoef, 2012). The p-values indicated that for the measurement of either μL or % volume change, FreeSurfer had better reproducibility than both manual (both $p \ll 0.001$) and FIRST (both $p \ll 0.001$), while the reproducibility of the manual method did not differ from that of FIRST (both $p > 0.5$).

Finally, our sample size ratio calculations indicate that, based on the reproducibility values and assuming all other effects constant, FreeSurfer would generally require the smallest sample sizes. Based on microliter volume change, and including failed segmentations,

Table 2
Median and interquartile range of longitudinal hippocampal volume change, averaged over back-to-back scans and left and right hippocampi. Data presented as μL change (upper block of rows) and % change from baseline (lower block of rows), for each of the three methods FIRST (left block of columns), manual outlining (middle block of columns) and FreeSurfer (right block of columns).

		FIRST		Manual		FreeSurfer	
	Subject group	Median	Interquartile range	Median	Interquartile range	Median	Interquartile range
Hippocampal volume change in μL	CTR	− 36	− 144 to + 46	− 103	− 199 to − 14	− 44	− 118 to + 21
	MCI-N	− 94	− 171 to + 23	− 109	− 181 to − 34	− 80	− 155 to + 8
	MCI-P	− 106	− 177 to + 2	− 92	− 202 to − 42	− 102	− 174 to − 15
	AD	− 96	− 233 to − 32	− 120	− 202 to − 48	− 105	− 198 to − 44
Hippocampal volume change in % of baseline volume	CTR	− 1.3	− 3.9 to + 1.2	− 3.0	− 5.6 to − 0.4	− 1.5	− 3.3 to + 0.7
	MCI-N	− 2.7	− 5.0 to + 0.7	− 3.5	− 5.8 to − 1.0	− 2.5	− 4.9 to + 0.3
	MCI-P	− 3.5	− 5.5 to + 0.1	− 3.3	− 6.9 to − 1.4	− 3.4	− 6.3 to − 0.6
	AD	− 3.2	− 7.6 to − 1.0	− 4.1	− 7.4 to − 2.0	− 4.2	− 7.6 to − 1.6

Table 3
Limits of agreement (LoA) for longitudinal data.

	LoA volume change μL (all data)	LoA volume change μL (failed segmentations removed)[a]	LoA % volume change (all data)	LoA % volume change (failed segmentations removed)[a]
FIRST	±478	±333	±18.8	±10.0
Manual	±313	±319	±9.5	±9.7
FreeSurfer	±313	±218	±23.1	±7.2

LoA = Limits of agreement, calculated by 0 ± 1.96 ∗ SEM√2 from linear mixed model analysis.
[a] Failed segmentations treated as individual missing data points in linear mixed model analysis.

FreeSurfer would require samples 7.4% smaller than FIRST, and 32.7% smaller than manual. Percentage volume change showed results contrary to this, due to outlier influence on these sample size calculations: FreeSurfer would require samples 72.6% larger than FIRST, and 46.9% larger than manual. Results for the comparison between FIRST and manual showed that FIRST required 27.3% smaller samples based on microliter volume change, and 48.5% smaller samples based on percentage volume change. When excluding failed segmentations, sample size reductions for FreeSurfer compared to manual were 57.8% (microliter volume change) and 50.7% (percentage volume change); for FreeSurfer compared to FIRST 38.8% (microliter) and 30.1% (percentage); for FIRST compared to manual 31.1% (microliter) and 29.5% (percentage).

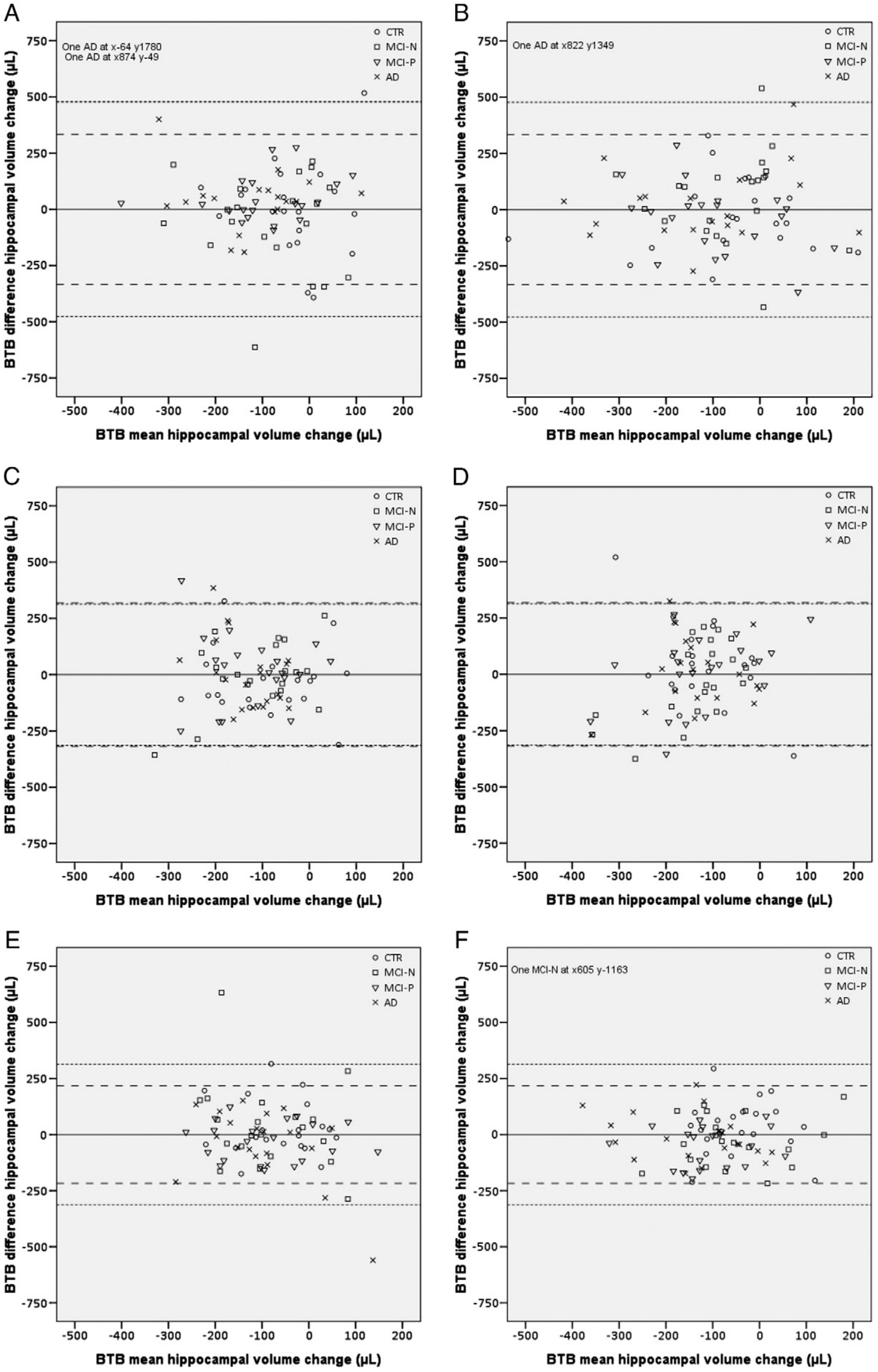
Discussion

This study quantitatively assessed the reproducibility of within-subject 1-year hippocampal volume change measurement by three methods: expert manual outlining, FIRST and FreeSurfer. While the reproducibility values, for both percentage and microliter hippocampal volume change, were roughly similar between the three methods, FreeSurfer did exhibit statistically significantly better reproducibility than manual or FIRST. In addition, while segmentation was incorrect in up to 7.5% of cases for the automated methods, there were no failed segmentations for the manual method.

Attenuation of disease progression is an important goal of treatment development for patients with MCI. Measurement of hippocampal atrophy rates in clinical trials can provide objective quantification of the ability of a new treatment to achieve this goal. The measurement methods used should not only be accurate, but also reproducible: by minimizing the uncertainty associated with the measurement method, the power to detect a treatment effect is maximized. Although inter-patient variability and other factors also affect power, better reproducibility of the measurement method can be expected to lead to smaller numbers of patients required per treatment arm, and hence reduced cost. Early work by Kikinis et al. (1992) on larger brain structures suggested that the variability of manual outlines of a brain structure increased with greater complexity and with smaller size, both of which would affect manual segmentation of the small and complex-shaped hippocampi negatively. It has been suggested by some (Dewey et al., 2010; Doring et al., 2011; Duchesne et al., 2002; Kennedy et al., 2009), and appears to be generally assumed, that due to the involvement of human operators, manual measurement of hippocampal atrophy rate would be less reproducible than automated methods. The results of the current study in part confirm this assumption, given the fact that reproducibility was significantly better for FreeSurfer compared to both manual measurement and FIRST, for both volumetric (μL) and relative (%) change. Nevertheless, reproducibility of hippocampal atrophy rates was similar between the manual method and both automated methods, with roughly comparable values, also for both μL and % change. It is possible that other automated methods than the two investigated here may yield smaller reproducibility values providing an even clearer separation from manual measurement. The sample size effect calculations based on the same joint linear mixed model analyses generally confirmed the superiority of FreeSurfer reproducibility, by showing the smallest required sample sizes for FreeSurfer except in case of severe outliers, while FIRST consistently required smaller sample sizes than the manual method. It should be noted that these sample size calculations are based solely on the reproducibility values for each method, and do not take into account other potentially important factors such as the difference between the clinical groups in the measured hippocampal volume change; hence they should be interpreted with some caution.

Manual measurement of hippocampal volume or atrophy rate is generally assumed to be more accurate than automated methods (Barnes et al., 2008; Boccardi et al., 2011). Further, manual measurement is generally used as the "gold standard" against which automated methods are validated in terms of accuracy (Dewey et al., 2010; Doring et al., 2011; Hsu et al., 2002; Kim et al., 2012; Lehmann et al., 2010; Morey et al., 2009; Pardoe et al., 2009; Sanchez-Benavides et al., 2010; Tae et al., 2008). We did not study accuracy of the methods in the current study; moreover, attempts to test that hypothesis may quickly lead to circular reasoning, since no real ground truth beyond that obtained by manual outlining is generally known. However, it seems to be a reasonable assumption that a well-trained operator would be able to delineate the hippocampus at least as well as these two widely used automated methods; one reason for this is that like most methods they are based on atlases or training data with manually delineated hippocampi as their core information (Fischl et al., 2002, 2004; Leung et al., 2010; Patenaude et al., 2011). Our observation of similar reproducibility, combined with this supposed superior accuracy for manual measurement, implies that the methodological design of large studies and clinical trials may favor manual measurement over these two widely used automated methods. There is a risk that for some methods, reproducibility may be good despite their inability to correctly outline the hippocampus. Moreover, due to the constraints placed on the models employed in such methods, the resulting volumes and shapes may be entirely plausible for hippocampus, even if they do not coincide with the hippocampus visible in the image. The rigorous quality control procedure with visual inspection of all segmentation outputs has ensured that such errors could not occur in the current study. Future studies should investigate the precise origins of segmentation errors, their anatomical distributions and their effects on change measurements, in order ultimately to devise improved methods.

Although it has been generally assumed that manual measurement would exhibit poorer reproducibility, this has not previously been tested directly. One practical reason for the absence of such studies is that the data necessary to perform them are generally not available. Manual segmentations vary due to operator expertise, operator fatigue and similar influences, and are reasonably expected to vary even when the imaging data does not change, as demonstrated among others by Warfield et al. (2004). Therefore, manual measurement reproducibility may be evaluated by asking an expert to process the same scan twice. However, the same approach would fail in the case of most automated methods, as they would simply produce exactly identical results, as has recently been demonstrated for FreeSurfer by Gronenschild et al. (2012) who demonstrated identical results both when identical analyses were performed one after the other, or in parallel on the same machine. Instead, in order to study reproducibility of a fully automated method, scan–rescan scan pairs appear to be the only viable approach, but


A
One AD at x-64 y1780
One AD at x874 y-49
B
One AD at x822 y1349
C
D
E
F
One MCI-N at x605 y-1163
○ CTR
□ MCI-N
▽ MCI-P
× AD
BTB difference hippocampal volume change (µL)
BTB mean hippocampal volume change (µL)
750
500
250
0
-250
-500
-750
-500
-400
-300
-200
-100
0
100
200

this necessarily implies an inability to disentangle acquisition-related and analysis-related influences. To determine in this way the reproducibility of a measure of change over time, scan–rescan scan pairs are required for two timepoints. Finally, in order to allow any conclusions that would generalize to a typical clinical trial setting, those images would have to be acquired at a number of different centers. ADNI provides a unique dataset that fulfills precisely these criteria: the within-session scan–rescan 3DT1 scan pairs ("back-to-back" scans) obtained at each timepoint allowed us for the first time to perform a direct, objective, quantitative comparison of reproducibilities between manual and automated measurements of hippocampal atrophy rates. A last issue is the substantial amount of work involved in the manual measurement. In order to measure the reproducibility of hippocampal atrophy rates in just 80 subjects in the current study, a total of 640 hippocampi had to be manually delineated on each slice (each subject had two timepoints, with two scans per timepoint, and two hippocampi).

By using joint linear mixed model analyses, we were able to perform direct comparisons between the three methods within the same model, thus avoiding any uncertainty that may arise from different model fit quality metrics in the case of separate models per method. Moreover, the linear mixed model analysis allowed handling of nested dependencies such as each patient having two hippocampi, modeling of both fixed and random effects variables, and handling of missing data, by which it provided superior estimates of each method's limits of agreement than those that could have been obtained from simpler analysis types (Bland and Altman, 1986, 1999, 2007). The similarity between the three methods in terms of LoA, observed for both volumetric and percentage change, was confirmed further by the median absolute differences. Although frequently reported in the context of methods evaluation, the median absolute difference is of less interest in the context of clinical trials because it does not reflect the tails of the distribution. Nevertheless this result confirms that manual measurement performs similarly to FreeSurfer and FIRST.

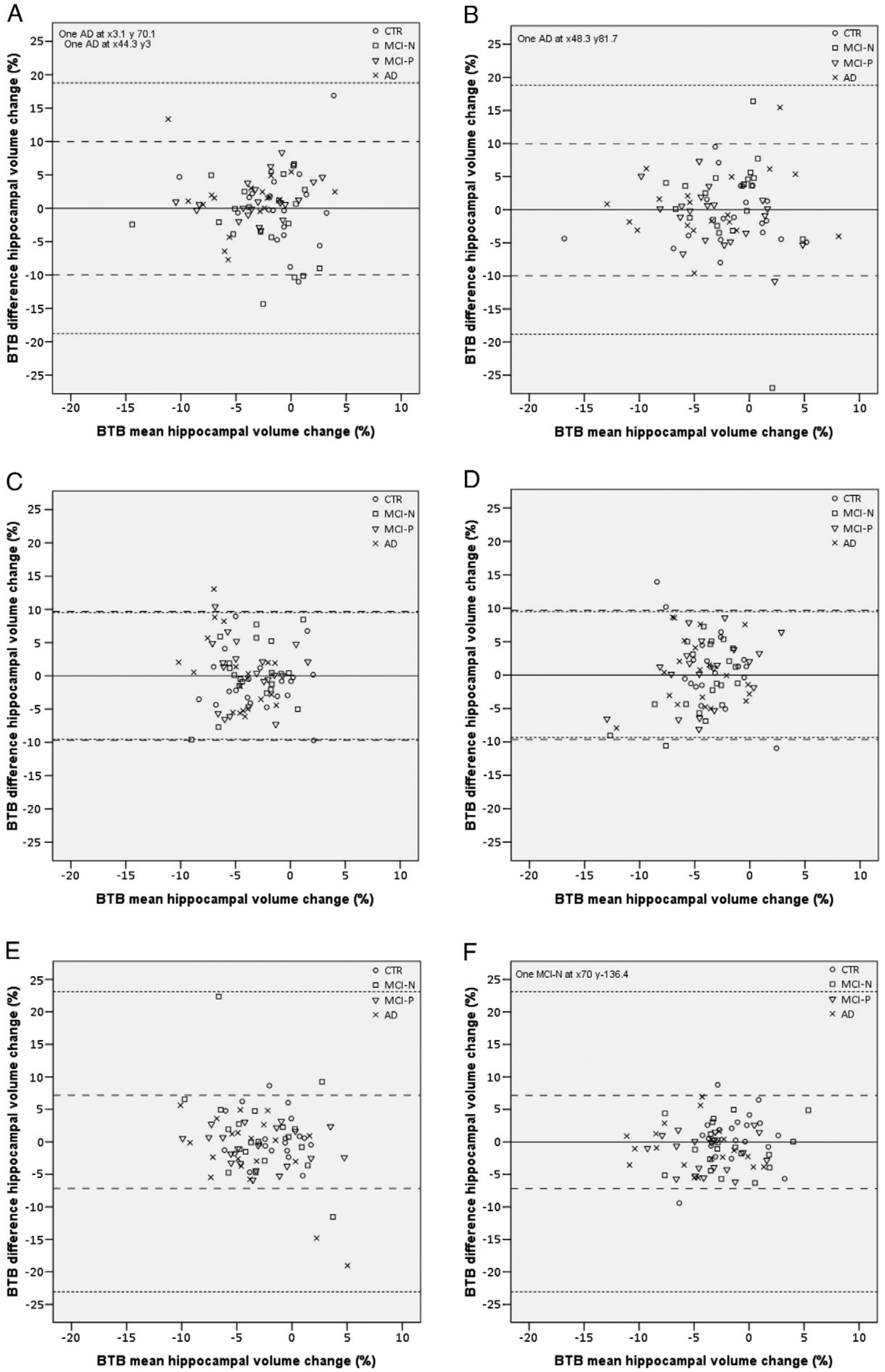
As larger and larger datasets are being used in clinical studies and trials, the need for increased automation also grows. The fully automated methods FreeSurfer and FIRST included in the current study fulfill this requirement to a substantial degree. Importantly, both FIRST and FreeSurfer segment subcortical gray matter structures in addition to the hippocampi, thereby delivering much more information than the manual method which just yields the hippocampal segmentations. A strong point of FIRST is its speed. FreeSurfer, quantifying cortical thickness in addition to the subcortical segmentation, requires considerably larger amounts of time, but this should not be a problem if sufficient computing power is available. However, both methods still require visual inspection of the output by a trained expert, whose training and work involve additional costs; moreover, this inspection leads to rejection of some segmentations. The failed segmentations occurred slightly more frequently in patients with AD than in MCI or healthy control subjects, but with just 80 subjects in total one should not draw firm conclusions from these findings. Our rather limited investigation did not reveal clear possible causes for the failed segmentations by FreeSurfer and FIRST in the current study, leaving full inspection of all output necessary. If such segmentation failure is discovered too late for the imaging to be repeated, this may lead to exclusion of that timepoint or even of the patient, from the study. In order to provide a fair comparison of costs, such exclusions based on failed automated segmentations should also be taken into account when calculating statistical power and study sample sizes. The expected rejection rate due to failed automated segmentations may be compensated for prospectively, by including additional patients, with associated additional costs.

In the present study we used, for any given hippocampus, all measurements from correct segmentations, even if the measurement of the same hippocampus by another method had failed. Exclusion of all measurements for that hippocampus, or of both hippocampi for that subject, would have led to exclusion of a large proportion of the subjects in the current study. Therefore, we have chosen to treat the failed segmentations as missing data points. The linear mixed model models these missing data points, allowing us to perform a fair comparison between methods without having to eliminate a large proportion of the data, and hereby draw valid conclusions on the comparison between the three methods.

Both FIRST and FreeSurfer were applied in this study using default settings. Specifically, for FIRST this implies a restriction of the number of modes of variation included in the modeling to 30 for each hippocampus. It cannot be excluded that in some individual cases the accuracy of the hippocampal segmentation by FIRST may be improved by including more modes of variation, but at a group level only minor improvement can be expected, as Patenaude and colleagues have demonstrated using repeated leave-one-out cross-validations for different numbers of modes of variation (Patenaude et al., 2011). For FreeSurfer, improved performance can be expected when applying manual editing. The full FreeSurfer pipeline recommends visual inspection and manual editing to correct errors at several stages of the process (http://surfer.nmr.mgh.harvard.edu/). This requires training and the editing itself takes time to perform, both of which lead to additional costs, although probably substantially less than the costs of fully manual segmentation. As a result, in large studies researchers often choose to accept the errors of the fully automated processing and use FreeSurfer as a fully automated method (e.g., Lucarelli et al., 2013; Westman et al., 2011). Therefore, in the current study we also used only the fully automated FreeSurfer pipeline to measure hippocampal atrophy rates. This also allowed us to compare two extreme approaches: fully manual segmentation on the one hand, versus fully automated segmentation on the other. Nevertheless, applying manual editing to the FreeSurfer or FIRST results should be expected to improve their performance, including an expected reduction of the number of "failed" segmentations.

The present work focuses on the rate of hippocampal atrophy, because this is the most important hippocampal outcome measure in the context of clinical treatment trials. Nevertheless, correct identification of the hippocampus on a single cross-sectional scan is still important. In cross-sectional validation studies across different diseases, FreeSurfer has been found to overestimate hippocampal volume when compared to manual delineation (Tae et al., 2008), especially in hippocampi that are subject to atrophy (Dewey et al., 2010; Kim et al., 2012; Lehmann et al., 2010; Pardoe et al., 2009; Sanchez-Benavides et al., 2010). FreeSurfer displayed better agreement with manual outlining than did FIRST (Doring et al., 2011; Morey et al., 2009; Pardoe et al., 2009), and similar distinctions between patient groups relative to control groups were reported for FreeSurfer compared to manual measurement (Lehmann et al., 2010; Shen et al., 2010). The present study adds substantially to the existing literature by demonstrating that the hippocampal atrophy rate, a more relevant measure than hippocampal volume, can be measured with roughly similar precision using FreeSurfer, FIRST, or manual measurement, with FreeSurfer exhibiting statistically significantly superior reproducibility. Moreover, the reproducibilities of the three methods were analyzed in a direct comparison, using a dataset that is very representative of those obtained in clinical trials.

Fig. 6. Modified Bland–Altman mean-difference plots for longitudinal hippocampal volume change (μL) between the two BTB scan pairs as measured by FIRST, manual segmentation and FreeSurfer, for left and right hippocampi. Dotted lines represent the limits of agreement (LoA) from the statistical model with all data included; dashed lines represent the LoA from the statistical model when failed segmentations are excluded from the model. Panels show results for: (A) FIRST, left hemisphere; (B) FIRST, right hemisphere; (C) manual segmentation, left hemisphere; (D) manual segmentation, right hemisphere; (E) FreeSurfer, left hemisphere; (F) FreeSurfer, right hemisphere. Abbreviations: CTR: healthy control; MCI-N: mild cognitive impaired with AD-negative t-tau/Aβ profile; MCI-P: mild cognitive impaired with AD-positive t-tau/Aβ profile; AD: Alzheimer's disease.


A
One AD at x3.1 y 70.1
One AD at x44.3 y3
○ CTR
□ MCI-N
▽ MCI-P
× AD
BTB difference hippocampal volume change (%)
25
20
15
10
5
0
-5
-10
-15
-20
-25
-20
-15
-10
-5
0
5
10
BTB mean hippocampal volume change (%)
B
One AD at x48.3 y81.7
○ CTR
□ MCI-N
▽ MCI-P
× AD
BTB difference hippocampal volume change (%)
BTB mean hippocampal volume change (%)
C
○ CTR
□ MCI-N
▽ MCI-P
× AD
BTB difference hippocampal volume change (%)
BTB mean hippocampal volume change (%)
D
○ CTR
□ MCI-N
▽ MCI-P
× AD
BTB difference hippocampal volume change (%)
BTB mean hippocampal volume change (%)
E
○ CTR
□ MCI-N
▽ MCI-P
× AD
BTB difference hippocampal volume change (%)
BTB mean hippocampal volume change (%)
F
One MCI-N at x70 y-136.4
○ CTR
□ MCI-N
▽ MCI-P
× AD
BTB difference hippocampal volume change (%)
BTB mean hippocampal volume change (%)

A limitation of the current study is its restriction to one manual method and two automated methods. For manual measurement, several protocols have been published and recently efforts have begun to harmonize protocols across the field (Boccardi et al., 2011). Each manual protocol is likely to have its specific advantages and disadvantages. The protocol used in the current study requires reformatting of the image volume perpendicular to the axis of the left hippocampus, which allows the expert operators to perform delineations always using more or less the same view, irrespective of the angulation of the head in the original, native image volume. This step however does require image intensity interpolation to the new voxel grid, which may lead to some loss of information. Secondly, the current manual protocol requires the follow-up scan to be analyzed side-by-side with the baseline scan. Since the order of the scans is therefore known to the operator, it cannot be excluded that this may in some cases lead to inflated atrophy rates. This may in part explain why the atrophy rates observed in the healthy control group were relatively high, when compared to values in the literature. For example, Jack et al. (1998) reported annual hippocampal volume loss of 1.55% (SD 1.38%) in healthy controls, compared to a mean annual loss of 3.1% (SD 3.8%) in the current study. Since no ground truth data are known, this issue cannot be resolved and may simply reflect small differences between study populations; e.g., some of the healthy controls in the current study converted to MCI. Further studies would be needed to determine whether the current manual method can lead to inflated atrophy rates. The design of future harmonized protocols for manual measurement of hippocampal atrophy rates should take such data into account (Boccardi et al., 2011). A recent paper by Maltbie et al. (2012) suggested that a large bias exists between manual outlines of the left and right hippocampus. This bias depended on the way in which images were presented to raters (either in their original orientation or in left–right mirrored fashion). It is unclear whether this bias would also influence the variability of the manual outlines differently for left and right hippocampi in our study. The linear mixed model analysis used included the different variance components involving hemisphere (i.e., H, HG, HM, HGM, PH:G, PHM:G, SH:PG, and SHM:PG), thereby explicitly modeling the possibility that the measurement-induced variance may differ between hemispheres. The limits of agreement therefore represent the total method-specific measurement uncertainty, incorporating any such effects if they exist. As argued by Maltbie and colleagues, since many automated methods rely on manual atlases as their core information, some of this asymmetry could be present in automated methods as well.

In addition to the two frequently used automated methods FreeSurfer and FIRST that were included in this study, a number of other fully automated methods have been described, including cross-sectional methods and methods designed to measure rates of atrophy (Barnes et al., 2004, 2007; Crum et al., 2001; Shen and Davatzikos, 2004; van der Lijn et al., 2008; Wang et al., 2003). Future studies should evaluate the performance of these methods using similar methodology to the present study, specifically using scan–rescan scan pairs to assess their reproducibilities. It appears likely that these advanced methods may also yield superior reproducibility. Furthermore, as fully automated methods may still be unsatisfactory and fully manual methods may be considered too costly with very large datasets, the performance of hybrid methods combining some expert input with automated processing should also be investigated. One example of this is the method previously described and validated by Crum et al. (2001) and van de Pol et al. (2007), which quantifies, using Jacobian integration of nonlinear image registration results, the total volumetric change in a region outlined by an expert on a single timepoint. By repeating the manual outlining of the hippocampi and registration procedures, van de Pol and colleagues found this hybrid method to have better reproducibility for atrophy rate measurement than the purely manual approach (Crum et al., 2001; van de Pol et al., 2007). Such hybrid approaches, including manual editing of FreeSurfer or FIRST results, should be investigated in carefully designed studies to determine the optimal combination of expert intervention and automated efficiency.

Finally, the “back-to-back” scans in this study were used without any correction beyond scanner corrections. Hence, geometric distortion due to gradient inhomogeneity was not fully corrected for. For whole-brain volume change measurement using SIENA, substantial effects of gradient-induced geometric distortions have been reported (Caramanos et al., 2010; Takao et al., 2010). Although because of the central location of the hippocampus in the brain, these effects are probably small in our case, they are present. This implies that some caution is warranted with respect to the volume changes measured by the three methods, but not with respect to their reproducibilities. Using rigid-body rotation between each back-to-back scan pair, we calculated the maximum displacement that had occurred anywhere inside the brain during the few seconds to minutes between the two acquisitions. For the 80 subjects in this study, this never exceeded 5 mm. Hence, the difference between the geometric distortions of the two scans of a single (within-session) back-to-back scan pair was small, allowing us to use them to assess reproducibility in a valid way.

Conclusion

Quantitative reproducibility values of 1-year microliter and percentage hippocampal volume change were roughly similar between expert manual outlining, FIRST and FreeSurfer, but FreeSurfer reproducibility was statistically significantly superior to both manual outlining and FIRST after exclusion of failed segmentations. In addition to reproducibility, hippocampal atrophy method related aspects that should be considered include accuracy, failing of several percent of automated segmentations, and the cost and training involved in visual inspection of segmentation results.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.01.058.

Acknowledgments

The authors thank Felix C. van Dommelen of the Image Analysis Center, VU University Medical Center, Amsterdam, The Netherlands for performing the manual hippocampal volume analyses, and Margo A. Pronk, of the same Image Analysis Center, for assistance in the visual inspection of segmentation outputs.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; Novartis Pharmaceuticals Corporation; Pfizer

Fig. 7. Modified Bland–Altman mean-difference plots for percentage longitudinal hippocampal volume change between the two BTB scan pairs as measured by FIRST, manual segmentation and FreeSurfer, for left and right hippocampi. Dotted lines represent the limits of agreement (LoA) from the statistical model with all data included; dashed lines represent the LoA from the statistical model when failed segmentations are excluded from the model. Panels show results for: (A) FIRST, left hemisphere; (B) FIRST, right hemisphere; (C) manual segmentation, left hemisphere; (D) manual segmentation, right hemisphere; (E) FreeSurfer, left hemisphere; (F) FreeSurfer, right hemisphere. Abbreviations: CTR: healthy control; MCI-N: mild cognitive impaired with AD-negative t-tau/Aβ profile; MCI-P: mild cognitive impaired with AD-positive t-tau/Aβ profile; AD: Alzheimer's disease.

Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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 Rev March 26, 2012 (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

Conflicts of interest

The Image Analysis Center, VU University Medical Center Amsterdam, a contract research organization, financed the manual hippocampal measurements by making available for this study one of their trained hippocampal measurement experts. Further, they employed E.R.M., R.A. de J. and R.A. van S. at the time of the study, and F.B. is their Director (as also indicated below).

E.R.M. was an employee of the Image Analysis Center, VU University Medical CenterAmsterdam, at the time of the study.

R.A. de J. is an employee of the Image Analysis Center, VU University Medical CenterAmsterdam.

D.L.K. has no conflict of interest.

R.A. van S. is partly working for the Image Analysis Center, VU Univerity Medical Center Amsterdam and partly working on the "neuGRID4you" project. The "neuGRID4you" project has received funding from the European Commission's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 283562.

K.S.C. is partly funded by the "neuGRID4you" project. The "neuGRID4you" project is funded from the European Commission's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 283562.

P.J.V. has served as an advisory board member of Bristol-Myers Squibb and Roche diagnostics. He receives/received research grants from Bristol-Myers Squibb, European Commission 6th and 7th Framework programmes, Joint Programme Initiative on Neurodegeneration, Zon-Mw, the Innovative Medicine Initiative, and Diagenic, Norway.

F.B. is Director of the Image Analysis Center, VU University Medical Center Amsterdam, and is a consultant for GE, Janssen Alzheimer Immunotherapy, and Roche Pharmaceuticals.

H.V. has received research grants from Pfizer, Novartis and Merck Serono, and speaker honoraria from Novartis, all paid to his institution.

References

Ard, M.C., Edland, S.D., 2011. Power calculations for clinical trials in Alzheimer's disease. J. Alzheimers Dis. 26 (Suppl. 3), 369–377.

Barnes, J., Scahill, R.I., Boyes, R.G., Frost, C., Lewis, E.B., Rossor, C.L., Rossor, M.N., Fox, N.C., 2004. Differentiating AD from aging using semiautomated measurement of hippocampal atrophy rates. Neuroimage 23, 574–581.

Barnes, J., Boyes, R.G., Lewis, E.B., Schott, J.M., Frost, C., Scahill, R.I., Fox, N.C., 2007. Automatic calculation of hippocampal atrophy rates using a hippocampal template and the boundary shift integral. Neurobiol. Aging 28, 1657–1663.

Barnes, J., Foster, J., Boyes, R.G., Pepple, T., Moore, E.K., Schott, J.M., Frost, C., Scahill, R.I., Fox, N.C., 2008. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. Neuroimage 40, 1655–1671.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.

Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. Stat. Methods Med. Res. 8, 135–160.

Bland, J.M., Altman, D.G., 2007. Agreement between methods of measurement with multiple observations per individual. J. Biopharm. Stat. 17, 571–582.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J.G., de Leon, M.J., deToledo-Morrell, L., Killiany, R.J., Lehericy, S., Pantel, J., Pruessner, J.C., Soininen, H., Watson, C., Duchesne, S., Jack Jr., C.R., Frisoni, G.B., 2011. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J. Alzheimers Dis. 26 (Suppl. 3), 61–75.

Caramanos, Z., Fonov, V.S., Francis, S.J., Narayanan, S., Pike, G.B., Collins, D.L., Arnold, D.L., 2010. Gradient distortions in MRI: characterizing and correcting for their effects on SIENA-generated measures of brain volume change. Neuroimage 49, 1601–1611.

Cover, K.S., van Schijndel, R.A., van Dijk, B.W., Redolfi, A., Knol, D.L., Frisoni, G.B., Barkhof, F., Vrenken, H., 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.

Crum, W.R., Scahill, R.I., Fox, N.C., 2001. Automated hippocampal segmentation by regional fluid registration of serial MRI: validation and application in Alzheimer's disease. Neuroimage 13, 847–855.

Dewey, J., Hana, G., Russell, T., Price, J., McCaffrey, D., Harezlak, J., Sem, E., Anyanwu, J.C., Guttmann, C.R., Navia, B., Cohen, R., Tate, D.F., 2010. Reliability and validity of MRI-based automated volumetry software relative to auto-assisted manual measurement of subcortical structures in HIV-infected patients from a multisite study. Neuroimage 51, 1334–1344.

Doring, T.M., Kubo, T.T., Cruz Jr., L.C., Juruena, M.F., Fainberg, J., Domingues, R.C., Gasparetto, E.L., 2011. Evaluation of hippocampal volume based on MR imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. J. Magn. Reson. Imaging 33, 565–572.

Drago, V., Babiloni, C., Bartres-Faz, D., Caroli, A., Bosch, B., Hensch, T., Didic, M., Klafki, H.W., Pievani, M., Jovicich, J., Venturi, L., Spitzer, P., Vecchio, F., Schoenknecht, P., Wiltfang, J., Redolfi, A., Forloni, G., Blin, O., Irving, E., Davis, C., Hardemark, H.G., Frisoni, G.B., 2011. Disease tracking markers for Alzheimer's disease at the prodromal (MCI) stage. J. Alzheimers Dis. 26 (Suppl. 3), 159–199.

Duchesne, S., Pruessner, J., Collins, D.L., 2002. Appearance-based segmentation of medial temporal lobe structures. Neuroimage 17, 515–531.

Euser, A.M., Dekker, F.W., le Cessie, S., 2008. A practical approach to Bland–Altman plots and variation coefficients for log transformed variables. J. Clin. Epidemiol. 61, 978–982.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der, K.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.

Fischl, B., van der, K.A., Destrieux, C., Halgren, E., Segonne, 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.

Gronenschild, E.H., Habets, P., Jacobs, H.I., Mengelers, R., Rozendaal, N., van, O.J., Marcelis, M., 2012. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS One 7, e38234.

Hoef, J.M.V., 2012. Who invented the delta method? Am. Stat. 66, 124–127.

Hsu, Y.Y., Schuff, N., Du, A.T., Mark, K., Zhu, X., Hardin, D., Weiner, M.W., 2002. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. J. Magn. Reson. Imaging 16, 305–310.

Jack Jr., C.R., Petersen, R.C., Xu, Y., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1998. Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. Neurology 51, 993–999.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27, 685–691.

Kennedy, K.M., Erickson, K.I., Rodrigue, K.M., Voss, M.W., Colcombe, S.J., Kramer, A.F., Acker, J.D., Raz, N., 2009. Age-related differences in regional brain volumes: a comparison of optimized voxel-based morphometry to manual volumetry. Neurobiol. Aging 30, 1657–1676.

Kikinis, R., Shenton, M.E., Gerig, G., Martin, J., Anderson, M., Metcalf, D., Guttmann, C.R., McCarley, R.W., Lorensen, W., Cline, H., 1992. Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging. J. Magn. Reson. Imaging 2, 619–629.

Kim, H., Chupin, M., Colliot, O., Bernhardt, B.C., Bernasconi, N., Bernasconi, A., 2012. Automatic hippocampal segmentation in temporal lobe epilepsy: impact of developmental abnormalities. Neuroimage 59, 3178–3186.

Lehmann, M., Douiri, A., Kim, L.G., Modat, M., Chan, D., Ourselin, S., Barnes, J., Fox, N.C., 2010. Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements. Neuroimage 49, 2264–2274.

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S., 2010. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 51, 1345–1359.

Lucarelli, R.T., Peshock, R.M., McColl, R., Hulsey, K., Ayers, C., Whittemore, A.R., King, K.S., 2013. MR imaging of hippocampal asymmetry at 3 T in a multiethnic, population-based sample: results from the Dallas Heart Study. AJNR Am. J. Neuroradiol. 34 (4), 752–757. http://dx.doi.org/10.3174/ajnr.A3308. (Electronic publication ahead of print 2012 Nov 8).

Maltbie, E., Bhatt, K., Paniagua, B., Smith, R.G., Graves, M.M., Mosconi, M.W., Peterson, S., White, S., Blocher, J., El-Sayed, M., Hazlett, H.C., Styner, M.A., 2012. Asymmetric bias in user guided segmentations of brain structures. Neuroimage 59, 1315–1323.

Morey, R.A., Petty, C.M., Xu, Y., Hayes, J.P., Wagner, H.R., Lewis, D.V., Labar, K.S., Styner, M., McCarthy, G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45, 855–866.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15, 869–877, xi-xii.

Pardoe, H.R., Pell, G.S., Abbott, D.F., Jackson, G.D., 2009. Hippocampal volume assessment in temporal lobe epilepsy: how good is automated segmentation? Epilepsia 50, 2586–2592.

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

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: a robust approach. Neuroimage 53, 1181–1196.

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61, 1402–1418.

Sanchez-Benavides, G., Gomez-Anson, B., Sainz, A., Vives, Y., Delfino, M., Pena-Casanova, J., 2010. Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer disease subjects. Psychiatry Res. 181, 219–225.

Schott, J.M., Bartlett, J.W., Barnes, J., Leung, K.K., Ourselin, S., Fox, N.C., 2010. Reduced sample sizes for atrophy outcomes in Alzheimer's disease trials: baseline adjustment. Neurobiol. Aging 31 (1452–62), 1462.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Shaw, L.M., Vanderstichele, H., Knapik-Czajka, M., Clark, C.M., Aisen, P.S., Petersen, R.C., Blennow, K., Soares, H., Simon, A., Lewczuk, P., Dean, R., Siemers, E., Potter, W., Lee, V.M., Trojanowski, J.Q., 2009. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Ann. Neurol. 65, 403–413.

Shen, D., Davatzikos, C., 2004. Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping. Neuroimage 21, 1508–1517.

Shen, L., Saykin, A.J., Kim, S., Firpi, H.A., West, J.D., Risacher, S.L., McDonald, B.C., McHugh, T.L., Wishart, H.A., Flashman, L.A., 2010. Comparison of manual and automated determination of hippocampal volumes in MCI and early AD. Brain Imaging Behav. 4, 86–95.

Tae, W.S., Kim, S.S., Lee, K.U., Nam, E.C., Kim, K.W., 2008. Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. Neuroradiology 50, 569–581.

Takao, H., Abe, O., Hayashi, N., Kabasawa, H., Ohtomo, K., 2010. Effects of gradient nonlinearity correction and intensity non-uniformity correction in longitudinal studies using structural image evaluation using normalization of atrophy (SIENA). J. Magn. Reson. Imaging 32, 489–492.

van de Pol, L.A., Barnes, J., Scahill, R.I., Frost, C., Lewis, E.B., Boyes, R.G., van Schijndel, R.A., Scheltens, P., Fox, N.C., Barkhof, F., 2007. Improved reliability of hippocampal atrophy rate measurement in mild cognitive impairment using fluid registration. Neuroimage 34, 1036–1041.

van der Lijn, F., den Heijer, T., Breteler, M.M., Niessen, W.J., 2008. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. Neuroimage 43, 708–720.

Wang, L., Swank, J.S., Glick, I.E., Gado, M.H., Miller, M.I., Morris, J.C., Csernansky, J.G., 2003. Changes in hippocampal volume and shape across time distinguish dementia of the Alzheimer type from healthy aging. Neuroimage 20, 667–682.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903–921.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., 2012. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement. 8, S1–S68.

Westman, E., Simmons, A., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Weiner, M.W., Lovestone, S., Spenger, C., Wahlund, L.O., 2011. AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. Neuroimage 58, 818–828.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation–maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.