Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease

Keith S Cover[a], Ronald A van Schijndel[a], Adriaan Versteeg[a], Kelvin K Leung[b], Emma R Mulder[a], Remko A Jong[a], Peter J Visser[a], Alberto Redolfi[c], Jerome Revillard[d], Baptiste Grenier[d], David Manset[d], Soheil Damangir[e], Paolo Bosco[c], Hugo Vrenken[a], Bob W van Dijk[a], Giovanni B Frisoni[c,f], and Frederik Barkhof[a], for the Alzheimer's Disease Neuroimaging Initiative*, neuGRID

[a]VU University Medical Center, Amsterdam, Netherlands [b]UCL Institute of Neurology, London, United Kingdom [c]IRCCS San Giovanni di Dio Fatebenefratelli, Italy, [d]MAAT, Archamps, France [e]Karolinska Institutet, Sweden, [f]University Hospitals and University of Geneva

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Corresponding Author

Keith S Cover, keith@kscover.ca, +31 65 596 7740
Physics and Medical Technology
VU University Medical Center
Amsterdam, Netherlands

***Highlights**

- Compared the reproducibility of methods for measuring hippocampal atrophy rates
- Compared FreeSurfer, FSL/FIRST, MAPS-HBSI, AdaBoost and manual
- Back-to-back MPRAGEs at baseline and year one for N=562 ADNI1 1.5T subjects
- Used a novel but simple statistical test that is robust to outlying data points
- MAPS-HBSI should require half the subjects in a clinical trial than others tested

Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease

Keith S Cover[a], Ronald A van Schijndel[a], Adriaan Versteeg[a], Kelvin K Leung[b], Emma R Mulder[a], Remko A Jong[a], Peter J Visser[a], Alberto Redolfi[c], Jerome Revillard[d], Baptiste Grenier[d], David Manset[d], Soheil Damangir[e], Paolo Bosco[c], Hugo Vrenken[a], Bob W van Dijk[a], Giovanni B Frisoni[c,f], and Frederik Barkhof[a], for the Alzheimer's Disease Neuroimaging Initiative*, neuGRID

[a]VU University Medical Center, Amsterdam, Netherlands [b]UCL Institute of Neurology, London, United Kingdom [c]IRCCS San Giovanni di Dio Fatebenefratelli, Italy, [d]MAAT, Archamps, France [e]Karolinska Institutet, Sweden, [f]University Hospitals and University of Geneva

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Corresponding Author

Keith S Cover, keith@kscover.ca, +31 65 596 7740
Physics and Medical Technology
VU University Medical Center
Amsterdam, Netherlands

Abstract

The purpose of this study is to assess the reproducibility of hippocampal atrophy rate measurements of commonly used fully-automated algorithms in Alzheimer disease (AD). The reproducibility of hippocampal atrophy rate for FSL/FIRST, AdaBoost, FreeSurfer, MAPS independently and MAPS combined with the boundary shift integral (MAPS-HBSI) were calculated. Back-to-back (BTB) 3D T1-weighted MPRAGE MRI from the Alzheimer's Disease Neuroimaging Initiative (ADNI1) study at baseline and year one were used. Analysis on 3 groups of subjects was performed – 562 subjects at 1.5T, a 75 subject group that also had manual segmentation and 111 subjects at 3T. A simple and novel statistical test based on the binomial distribution was used that handled outlying data points robustly. Median hippocampal atrophy rates were -1.1% /year for healthy controls, -3.0%/year for mildly cognitively impaired and -5.1% /year for AD subjects. The best reproducibility was observed for MAPS-HBSI (1.3%), while the other methods tested had reproducibilities at least 50% higher at 1.5T and 3T which was statistically signifigant. For a clinical trial, MAPS-HBSI should require less than half the subjects of the other methods tested. All methods had good accuracy versus manual segmentation. The MAPS-HBSI method has substantially better reproducibility than the other methods considered.

Keywords

Hippocampus, atrophy, manual segmentation, automatic segmentation, magnetic resonance imaging, mild cognitive impairment, Alzheimer disease, boundary shift integral

1. Introduction

A feature of Alzheimer's disease (AD) (Jack et al., 1992; Jack et al., 1998; Wang et al., 2003; Evans et al., 2010; Frisoni et al. 2010; Drago et al., 2011) is increased hippocampal volume loss when compared to age matched healthy controls (HC). Mildly cognitive impairment (MCI) subjects typically have intermediate hippocampal volumes and rates of loss. Hippocampal atrophy rates have been proposed (Schott et al., 2010; Ard and Edland, 2011) or used (Wilkinson et al. 2012) as end points in clinical trials. Manual segmentation of hippocampi (Barnes et al., 2008; Boccardi et al., 2011) is often regarded as the "gold standard" for volume measurement – however this may take about 3 hours per MRI scan (Mulder et al., 2014) and requires extensive training. The size of AD clinical trials (typically many hundreds of subjects) means that there is great interest in less labour-intensive methods; as a result several fully automated techniques have been developed and are increasingly used.

Manual measurements of hippocampal volume or atrophy rate are generally assumed to be more accurate than automated methods (Barnes et al., 2008; Boccardi et al., 2011) and are used for validation of the accuracy of automated techniques (Hsu et al., 2002; Tae et al., 2008; Morey et al., 2009; Pardoe et al., 2009; Dewey et al., 2010; Lehmann et al., 2010; Sanchez-Benavides et al., 2010; Doring et al., 2011; Kim et al., 2012; Iglesiasa et al., 2015). However, fully automatic methods have improved to the point where it has been suggested that they have similar accuracy when compared to manual measures and are more reproducible (Duchesne et al., 2002; Kennedy et al., 2009; Dewey et al., 2010; Doring et al., 2011). As a consequence, a number of comparisons of methods for measuring atrophy rates have been published (Kikinis et al., 1992; Fox et al. 1997; Rudick et al., 1999; Crum et al., 2001; Zhang et al., 2001; Smith et al., 2002; Barnes et al., 2004, 2007; van de Pol et al., 2007; Altman et al., 2009; Barkhof et al., 2009; Shaw et al., 2009; Sluimer et al., 2009; Shen et al, 2010; Westman et al., 2011).

The ideal way to compare atrophy rate measurement methods would use perfectly accurate segmentations as a gold standard. The performance of each method could then be compared against the perfect segmentation over a set of subjects. By calculating the spread of the errors in each method – such as the standard deviation – the best performing methods could be determined. Perfectly accurate segmentations are not available, but we can obtain an indication of the spread of the errors in the methods - provided the methods are reasonably accurate - by repeating the measurements and determining their spread.

The goal of the current study was to compare the reproducibility of hippocampal atrophy rate of commonly-used automated measurement techniques, at both 1.5T and 3T, taking advantage of back-to-back (BTB) MPRAGE volumetric scans routinely acquired at each subject in the first Alzheimer's Disease Neuroimaging Initiative (ADNI1) study. We aimed to assess the most recent versions of FreeSurfer (Fischl et al., 2002, 2004; Reuter et al., 2012), FSL/FIRST (Patenaude et al., 2011), AdaBoost (Morra et al., 2009) and MAPS-HBSI (Leung et al. 2010).

The data set from the ADNI1 study (Mueller et al., 2005; Jack et al., 2008; Weiner et al., 2012) provides a singular opportunity to compare the reproducibilities of brain atrophy methods. While rarely mentioned in the literature, as part of ADNI1, two 3D T1 weighted MPRAGEs were acquired BTB during each subject visit - with the acquisition of the second MPRAGE usually starting within seconds of completion of the first (Cover et al. 2011). All ADNI1 subjects were asked to have a scan at 1.5T with a subset of subjects also having 3T imaging. With 800 subjects acquired across 55 sites included in ADNI1, it provides a much larger BTB dataset than available for previous reproducibilities studies. In addition, the ADNI1 study put a great deal of effort into standardizing the acquisition of the MPRAGE sequences across the ADNI1 sites. Thus, ADNI1 provides an excellent dataset to test the reproducibility of the measurement of hippocampal atrophy rates and other structural segmentation methods.

For the hippocampus atrophy rates, the BTB reproducibility of manual segmentation at 1.5T of hippocampi atrophy (Mulder et al., 2014) has been compared to FreeSurfer, and FSL/FIRST for a subset of N=80 subjects of the ADNI1 dataset. Mulder et al. found the manual and automated segmentations had similar reproducibilities.

Although the ADNI1 study was performed primarily at 1.5T, with research studies and trials in AD and other disorders shifting to 3T acquisitions (de Jong et al., 2008; Watson et al. 2010) it was important to include in ADNI1 a sub-set of subjects who had 3T BTB as well as 1.5T BTB imaging. A direct comparison between 3T and 1.5T has only been performed for a cross sectional method (Keihaninejad et al. 2010) but without reproducibility measurements. Longitudinally, only the reproducibility of the FSL/Siena measure for whole brain atrophy has been compared at 1.5T and 3T (Cover et al. 2014).

In addition, for whole brain volume atrophy measures at 1.5T (Popescu et al. 2012), subsets of the ADNI1 BTB dataset have been used to compare the reproducibility (Cover et al., 2011) of Siena and SienaX .

Here, we compared the reproducibility of 7 popular methods to determine hippocampal atrophy rates over 1 year. Such information is important to plan clinical trials in AD.

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

assessment 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.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California–San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

2. **Materials and Methods**

2.1 ADNI1 dataset

In ADNI1 two BTB MPRAGEs were acquired with identical acquisition parameters during each of the two subject visits – baseline and year one – without removing the subject from the scanner (Jack et al., 2008). Referred to as "original" MPRAGEs by ADNI1, for the current study the first acquired original MPRAGE is referred to as "M" and the second as "N". ADNI selected one of M or N for additional processing and produced a third MPRAGE - referred to as "processed" by ADNI - for each subject visit. The processed MPRAGE is referred to as "P" in this study. The additional ADNI processing to generate P included B1 non-uniformity correction, intensity nonuniformity correction and gradient warp correction (Jack et al., 2008; Clarkson et al., 2009). Fig. 1 illustrates the relationship of the 6 MPRAGEs for each subject. While M and N provide information on reproducibility, P provides accuracy information used to ensure the atrophy rates of the methods are accurate enough that the reproducibilities are meaningful.

The M, N and P MPRAGEs used in the current study are exactly those downloaded from ADNI. According to ADNI, the M and N voxel values – which were 16 bit values - are unchanged from those generated by the MRI scanners. Only some of the meta data of the DICOM files were modified by ADNI. Before they were processed by the methods in the current study, no processing was applied to either of the M or N MPRAGEs, other than conversion from the DICOM to NIFTI file format. The same DICOM to NIFTI conversion was used for all methods and the 16 bit voxel values were unchanged in the conversion. Differing from M and N, the P voxel values as supplied by ADNI were 32 bit floating point and were converted to 16 bit integer as part of the study's conversion from DICOM to NIFTI.

A total of 4,038 MPRAGEs were included in the current study. At 1.5T there were 3,372 MPRAGEs. For each of the N=562 subjects at 1.5T, 6 MPRAGEs were included – 3 MPRAGEs for the baseline subject visit and 3 for the year one subject visit. Similarly at 3T, the 111 subjects had a total of 666 MPRAGEs.

The 562 subjects at 1.5T are a subset of the first year collection of 639 subjects from the ADN1 study (Wyman et al. 2013). Only subjects with exactly 2 MPRAGE acquired at both the baseline and year one subject visits were included in the 562 subjects analyzed in the current study. We also excluded subjects with 3 or more MPRAGE acquired at either the baseline or year one visit, since this might indicate that one of the first two MPRAGEs acquired had a serious problem. Even with those subjects excluded, the 562 subjects included in the study is a large set compared to previous reproducibility studies.

The manual segmentation of a subset of ADNI1 BTB MPRAGEs at 1.5T (Mulder et al. 2014) were used in the current study to determine the accuracy of the automated methods. While there were N=80 subjects in the Mulder et al. study only N=75 were also in the N=562 subject group of the current study. Therefore only N=75 subjects were included in the manual subset of the current study.

The ADNI1 study also acquired BTB MRPAGEs at 3T of a subset of subjects (Wyman et al. 2013). Of the N=562 included at 1.5T, a total of N=111 had exactly 2 MPRAGEs acquired at 3T at both the baseline and year one subject visits and were included in the current study.

The 562 subjects in the current study consisted of 171 HC, 277 MCI and 114 AD based on the ADNI1 classification of the subjects. The median age of the HC subjects was 75.7 (72.5, 78.7) - the numbers in brackets are the interquartile range of age -with 50% male. For MCI the median age was 75.2 (70.7, 79.9) with 65% male. And for AD the median age was 75.6 (70.3, 81.2) with 50% male.

For both 1.5T and 3T all pixels were square and the slice thickness was 1.2mm. For 1.5T the voxel volume ranged from 1.05mm^3 to 2.20mm^3 with a median value of 1.05mm^3. For 3T it ranged from 1.20mm^3 to 1.24mm^3 with a median of 1.20mm^3.

2.2 Hippocampal atrophy rate measurement

A total of seven methods for measuring hippocampi atrophy rates were included in the study – manual segmentation and six fully automatic methods: the automated methods had four "cross-sectional methods" and two longitudinal methods. For the purposes of this study we refer to cross-sectional methods as those that calculate a hippocampal volume for the MPRAGE at each time point and the volumes are then subtracted to derive a volume difference from which atrophy rates are calculated. Longitudinal methods analyze the two MPRAGEs at the two time points simultaneously with the aim of improving the precision of the atrophy rate calculations.

The four cross-sectional methods included in the current study were FMRIB's Integrated Registration and Segmentation Tool FSL/FIRST 5.0.4 (Patenaude et al., 2011) which is part of the FMRIB Software Library (FSL) (fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST). The second was AdaBoost a "machine learning" based segmentation method (Morra et al., 2009). The particular implementation of AdaBoost used in the current study was implemented by one of the authors (AR) and trained on the harmonized protocol for hippocampal segmentation (Frisoni et al. 2014) (neugrid4you.eu). The third was FreeSurfer/ReconAll 5.3.0 in cross sectional mode (FreeSurferC) (Fischl et al., 2002, 2004; Reuter et al., 2012) (http://surfer.nmr.mgh.harvard.edu). The fourth was the Multiple-Atlas Propagation and Segmentation algorithm (MAPS) (Leung et al., 2010) and was implemented by an author of the current paper (KKL) (neugrid4you.eu). MAPS was tuned and tested against the ADNI1 1.5T data set as part of it implementation. All pipelines were run in their default mode.

The 2 longitudinal algorithms included were extensions of two of the cross-sectional algorithms. They were FreeSurfer/ReconAll 5.3.0 in longitudinal mode (FreeSurferL) (http://surfer.nmr.mgh.harvard.edu) and the Multiple-Atlas Propagation and Segmentation with Hippocampal Boundary Shift Integral (MAPS-HBSI) (Leung et al., 2010; Barnes et al., 2007; Freeborough and Fox, 1997). An author of the current paper (KKL) was involved in implementing the HBSI component of MAPS-HBSI. While MAPS was tuned against the ADNI1 dataset the HBSI algorithm was not tuned using the ADNI1 data set or any other data set.

All automated hippocampal segmentations were performed on 64-bit Linux machines.

The manual segmentation included in the current study used the results of the segmentation performed for Mulder et al. 2014. The details of the manual segmentation are described in that paper. While Mulder et al. included N=80 subjects only N=75 of those subjects were used in the current study as only 75 of the subjects were included in the ADNI1 collection used in the current study. The 75 subject subset included 19 HC, 38 MCI and 18 AD subjects.

2.3 Statistics

Calculation of atrophy rates from the hippocampal volumes at baseline (V_A) and year one (V_B) is straightforward. The non-annualized percentage volume change (PVC) was calculated by $100*(V_B – V_A)/V_A$. For each subject there was a PVC calculated for each of the M, N and P scan pairs. The PVC can be annualized by adjusting for the exact interval between baseline and one year scans.

The BTB difference is defined as PVC_N-PVC_M . The BTB difference provides a measure of the reproducibility of the non annualized PVC for each subject. The non annualized PVC is used for calculating the BTB difference as the BTB difference appears to be independent of the interval between the scans (Cover et al. 2014). The closer to zero the BTB difference the more reproducible the atrophy rate measure. Of course, for the BTB difference to be meaningful, the algorithm must be accurate.

It has been previously observed (Cover et al. 2014) that BTB differences for whole brain atrophy rates have a large number of outlying points, making standard statistical tests problematic. One of the two techniques used in the current study to compare the BTB difference of the various methods was presented by Smith et al. (2007). For each method they calculated the median of the magnitude of the BTB differences. The median – which will also be referred to as the method reproducibility in the current paper - provides a measure of the width of the BTB differences for each method and allows the methods to be ordered by reproducibility.

Taking the magnitude eliminates the sign of each BTB difference before the median is taken. Thus the median of the magnitude of the BTB differences is a measure of the spread of the BTB differences of a method. The method used by Smith et al. is similar to a standard deviation but is less sensitive to large numbers of outlying BTB differences common in some atrophy measures.

However, one problem with the Smith et al. method is that it does not provide an easy way to determine if the difference between the median of two methods is statistically significant unless certain assumptions hold - such as the BTB differences for a method have a normal distribution. Cover et al. (2014) demonstrated that for whole brain atrophy measures BTB difference distribution has far too many outlying points to be treated as normal.

The second technique used in the current study to compare the BTB differences of the various methods is a simple statistical test that handles outlying statistical points robustly and is being introduced in the current study. Based on binomial statistics, it handles both normal and non-normal distributions accurately - including ones with large shoulders. Also, the binomial based test does not require the reproducibility to be the same across subjects and sites. For example, if the reproducibility happens to be poorer in subjects with more advanced AD, the binomial test will handle it properly.

When comparing two methods, the first step of the binomial test compares the magnitudes of the BTB differences on a subject-by-subject basis and calculates the fraction of the subjects that is larger for one method. Thus, if the fraction is 0.5 then the two methods are statistically equal. When the fraction is different from 0.5, the p-value to reject the null hypothesis can be calculated using the binomial distribution. Thus the null hypothesis is when the fraction is 0.5.

3. Results

Table 1 presents the method reproducibilities at 1.5T for each of the seven methods. For both left and right hippocampi, and also for N=562 and N=75, the MAPS-HBSI algorithm has the lowest method reproducibility – where the method reproducibility is the median absolute BTB difference for the method. The longitudinal mode of FreeSurfer is second best hippocampus for N=562 as its reproducibility is 62% larger - and thus worse - than that of MAPS-HBSI for the left hippocampus, and 54% for the right.

Fig. 2 shows scatter plots of the BTB differences of both FreeSurfer in longitudinal mode and manual versus MAPS-HBSI for the manual group (N=75). In both cases, the reproducibility distribution is smaller for MAPS-HBSI in agreement with Table 1.

Table 2 presents the fraction of subjects for each method at 1.5T where the magnitude of the BTB difference is larger than MAPS-HBSI. The comparison is limited to MAPS-HBSI as it was the method with the best reproducibility. For all methods, the fraction is greater than 0.5 - which is inline with MAPS-HBSI's having the better reproducibility. For N=562 the p-value is less than 0.00001 for all methods. The p-values for N=75 for manual compared to MAPS-HBSI were 0.0008 for left and 0.0237 for right – which are both statistically significant. Thus the results indicate MAPS-HBSI is more reproducible than manual at 1.5T.

Comparison of the accuracy of the fully automatic methods at 1.5T is given in Table 3. The table presents the annualized percentage volume changes – also called the atrophy rates - at 1.5T for P for all the methods with the exception of manual. It also presents the annualized percentage volume change for M and N for the two longitudinal methods - FreeSurfer in longitudinal mode and MAPS-HBSI. M and N were provided for the longitudinal algorithms for comparison. The atrophy rates are presented for HC, MCI and AD. Median hippocampal atrophy rates were similar for the different methods justifying the comparison of their reproducibilities. MAPS-HBSI for P yielding –1.1% /year for healthy controls, -3.0%/year for mildly cognitively impaired and –5.1% /year for AD subjects.

Comparison of the reproducibilities of FreeSurfer and MAPS-HBSI at 1.5T versus 3T is presented in Fig. 3 and Table 4. Fig. 3 shows scatter plots for both FreeSurfer (a) and MAPS-HBSI (b) of the reproducibility of each subject for 3T versus 1.5T for the left and right hippocampi. The larger scatter of FreeSurfer both horizontally (1.5T) and vertically (3T) demonstrates MAPS-HBSI has better reproducibility at both 1.5T and 3T. For the left hippocampus, the reproducibility of MAPS-HBSI is better than FreeSurfer at 3T ($p<0.00001$) by 87%. For the right hippocampus, the reproducibility of MAPS-HBSI is better than FreeSurfer at 3T ($p=0.00005$) by 66%.

Although the focus of the current study was the reproducibility of the methods it was important to determine whether the method with the best reproducibility, MAPS-HBSI, was accurate. Fig. 4 shows a Bland-Altman plot (Bland et al. 1986) of the annualized percentage volume change for manual versus MAPS-HBSI at 1.5T for the manual group of 75 subjects. As manual segmentation is considered to be the gold standard for accuracy, the Bland-Altman plot indicates MAPS-HBSI has good accuracy for measuring hippocampal atrophy rate and its method reproducibility can be considered the noise of the accuracy.

Table 5 presents the failure rate of the methods for 1.5T and 3T. A method is considered to have failed when its analysis of a MPRAGE or pair of MPRAGEs has not generated all the values for the required volumes and/or atrophy rates. Interestingly, the two longitudinal methods – FreeSurferL and MAPS-HBSI – have failure rates of about 4% for the ADNI processed MPRAGEs (P) at 1.5T as compared to about 1% for the original MPRAGEs (M and N) at both 1.5T and 3T. We examined the failed output in some detail but could not come up with any solid conclusion. Few methods failed on the same MPRAGE so the failures had little to do with bad MPRAGES. However, it is important to note that only 3 subjects failed both of the longitudinal methods. The rest only failed one of either FreeSurferL or MAPS-HBSI. Also, a patient that failed M or N was no more likely to fail P. That each method generally failed on a different MPRAGE suggests the failures had little to do with bad MPRAGES.

The run times for ADABoost and FSL/FIRST were one to two hours per MPRAGE. For FreeSurfer the run time per MPRAGE was about 24 hours in cross sectional mode. FreeSurfer in longitudinal mode requires about 72 hours per pair of MPRAGEs. However, the first 48 hours also generated the volumes for the cross sectional mode for each of the two MPRAGEs so the longitudinal mode is only an additional 24 hours of calculation over the cross sectional mode. MAPS-BHSI is about 96 hours for longitudinal mode but it also generates the MAPS cross sectional volumes for the two MPRAGEs. As mention above the manual segmentation time was about 3 hours per MPRAGE.

4. Discussion

The 562 subjects in ADNI1with BTB MPRAGEs acquired at 1.5T, along with a sub group of 111 subjects with BTB MPRAGEs also acquired at 3T, provide an invaluable data set based on true scan-rescan imaging that can be used to compare the reproducibility of methods to measure atrophy. In this study we focused on the hippocampal atrophy rate, an important end-point in AD trials. The results of the current study showed that MAPS-HBSI is substantially more reproducible than the other methods included in the current study including manual measurements – the gold standard in the field. We used the binomial statistical test that provides a robust and simple way to assess whether there is a significant difference in reproducibility among the various methods.

The better reproducibility of MAPS-HBSI (Fig. 4) suggests that for the same statistical power to detect a change in atrophy rates it would require substantially smaller sample sizes in studies than the other methods included in the current study. A simple group size calculation based on the square-root-of-N rule indicates MAPS-HBSI should require less than half the number of subjects to detects the same change of atrophy rate. For example, for N=562 the reproducibility of MAPS-HBSI and longitudinal FreeSurfer for the left hippocampus are 1.3% and 2.1%. The relative group size is then 2.6 (= (2.1%/1.3%)^2) by the square-root-of-N rule. Therefore FreeSurfer would require 2.6 times as many patients in a study as HBSI to detect the same change in the PVC . While the square-root-of-N rule makes assumptions that may not strictly hold for the BTB differences, it is likely at good approximation of the correct value.

As opposed to the other methods in the current study, HBSI is a step applied after hippocampal delineation and segmentation. As mentioned above, all the other methods segment the hippocampi with the smallest unit being discrete voxels. Thus there is nothing to prevent HBSI from being applied as an additional step after the segmentations of the other algorithms. For example, Leung et al. 2010 also applied HBSI to manual segmentation.

MAPS-HBSI is unique among the methods considered in this study in that, while the other methods classify each voxel as either fully in or fully out of the hippocampus, MAPS-HBSI uses the intensity values of the voxels to take into account partial volume effects. Methods such as FSL/FIRST and FreeSurfer may use partial volumes during their calculations however both present their results as a mask with the same voxel size as their respective MPRAGEs with each voxel being assigned a zero or one. Thus such volumes do not take partial volumes into account. When partial volumes are taken into account, non hippocampal structures are less likely to be included in the hippocampus volume change calculations. While it is not currently clear if the partial voxel volume nature of HBSI is the key to its success, applying HBSI to the output of the other methods will provide valuable insight into this question.

As mentioned in the Methods section, MPRAGEs from the ADNI1 data set were used to tune the MAPS method. This raises the question of whether the ADNI1 tuning may have given MAPS-HBSI a special advantage in this study. Such an advantage is judged unlikely for two reasons. First, the MAPS method used without HBSI in the current study had reproducibility in line with the other methods and not as good as MAPS-HBSI, while the ADNI1 data set was not used to calibrate the HBSI step that followed MAPS. It was the HBSI step following MAPS that yielded the increase in reproducibility. The advantage MAPS-HBSI introduces over other methods is likely to apply to the segmentation of other methods and should be the focus of further research.

For several reasons no manual review of the automatic segmentation was performed in the current study even though it is often recommended. First, the primary goal of this study was to compare the performance of fully automatic measurements of hippocampal atrophy rates thus manual review of the segmentation was unnecessary. Second, with segmentations of 40,464 hippocampi required at 1.5T - both left and right, baseline and year one, and M, N and P for each of the 6 methods for N=562 - manual review of the segmentations was unattainable. Third, the accuracy of MAPS-HBSI was assessed by comparing its atrophy rates to those of manual segmentation – a comparison that shares some of the benefits of a manual review. In addition, while a standard for manual segmentation is coming closer to reality (Boccardi et al., 2011; Frisoni et al., 2014), it still needs to be widely accepted.

The reproducibility for the hippocampal atrophy rates for FreeSurfer in the current study was smaller – in other words better performing - than for Mulder et al. (2014). In both cases the reproducibility of a method was calculated by the median of the magnitude of the BTB differences – the same way the method reproducibility was calculated in the current paper. The left and right reproducibilities for N=75 for longitudinal FreeSurfer in the current study are 1.9% and 1.7% while Mulder et al. for N=80 found 2.5% and 2.5%. The lower values of the current study are likely due to a newer version of FreeSurfer - 5.3.0 for the current study compared to 5.1.0 for Mulder et al. It is unlikely the lower values are due to the 5 subjects not included in the manual group of the current study as all the subjects included for manual in the current study were used in Mulder et al. and outlines used in the current study are also the same as those used in Mulder et al. For manual segmentation, the current study had 2.62% (left) and 2.59% (right) (N=75) while the previous had 2.5% and 3.6% (N=80) indicating removing the 5 subjects had little impact on the results. Thus it is likely the improved reproducibility of FreeSurfer in the current study is due to the newer version of FreeSurfer but a more detailed analysis would be required to make any definitive statements.

For MAPS-HBSI in the current study the median annualized PVC for HC (N=171) was -1.4% (M) and -1.0% (N) for the left hippocampi and -1.3% (M) and -0.9% (N) for the right (Table 3). These values are in line with Jack et al. (1998) who reported an annualized PVC of -1.55% in healthy controls but on a different population.

A wide range of techniques and statistical methods have been used in the literature to compare the performance of methods for measuring hippocampal atrophy rates. A commonly used technique is based on predicting which subjects will convert from MCI to AD. However, the clinical differentiation between MCI and AD does not match well with the pathological classification. As reported by Schneider et al. (2009), about half the subjects classified MCI clinically have AD pathology. Another common assumption used that does not always hold is atrophy rate reproducibility has a normal distribution. Cover et al. (2014) demonstrated the reproducibility of whole brain atrophy does not have a normal distribution because of its many outliners. Both the statistical tests used in the current study handle outliers robustly.

The current paper introduces a statistical test – based on the binomial distribution - to determine if, for the same set of subjects, a set of BTB differences for one method is larger than a set for a second method and whether the difference is statistically significant. The main advantage of the test is it makes few assumptions about the set of BTB differences for a method. For example, the test

does not assume the BTB differences is normally distributed or that the reproducibility does not vary from subject to subject. For example, the BTB differences may increase, with more advanced disease for vary from site to site.

The high failure rates of about 4% for P for MAPS-HBSI and FreeSurfer as compared to about 1% for M and N could have a number of causes. One possibility may be due to both MAPS-HBSI and FreeSurfer performing their own distortion and inhomogeneity corrections. While M and N do not have the corrections, P does. Thus, the higher failure rate might be due to the distortion and inhomogeneity correction being applied twice. As 4% is a significant loss of subject data the cause of the failures deserves additional scrutiny.

As the gradient warp varies from site to site the accuracy of hippocampal volumes may be affected. However, the current study is interested in hippocampal atrophy rates not volumes. Since the gradient warp corrections are relatively small and should be nearly identical for both M and N - as the subject was left in the same position for both BTB MPRAGEs - gradient warp correction should have little impact on atrophy rate calculations (Caramanos et al, 2010; Takao et al., 2010).

5. Conclusions

The MAPS-HBSI algorithm is more reproducible when measuring hippocampal atrophy rates than the other methods included in the current study. Based on the results of the current study, in a clinical trial the other methods should require at least twice the subjects of MAPS-HBSI to detect the same change in atrophy rate. In addition, all methods tested had good accuracy versus manual segmentation.

Given the result that MAPS-HBSI has substantially better performance than the other methods in the current study, it may be worth including in studies where improved measurement of hippocampal atrophy rates may be of benefit, such as in clinical trials.

Acknowledgements

Thanks to Nick C Fox for help with the MAPS-HBSI algorithm and for valuable comments on the manuscript.

Study funding was provided by neuGRID4you (www.neuGRID4you.eu), an European Community FP7 project (grant agreement 283562), and the VU University Medical Center, Amsterdam, The Netherlands.

All the algorithms in this study are implemented on the neuGRID computer infrastructure (www.neugrid4you.eu) (Frisoni et al., 2011). Most of the jobs for this study were run on neuGRID and are available for academic use. Additional calculations were also performed on the NCAGRID cluster at the VU University Medical Center.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012) ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; 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.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics,LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; 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 (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 Southern California.
.

Conflicts of interest

The “neuGRID4you” project is funded from the European Commission's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 283562.
The Image Analysis Center, VU UniversityMedical Center Amsterdam, a contract research organization, financed themanual hippocampalmeasurements 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).
K.S.C. is partly funded by the “neuGRID4you” project.
R.A. van S. is partly working for the Image Analysis Center, VU Univerity Medical Center Amsterdam and partly working on the “neuGRID4you” project.
A.V. is partly funded by the “neuGRID4you” project.
E.R.M. was an employee of the Image Analysis Center, VU University Medical CenterAmsterdam,
R.A. de J. is an employee of the Image Analysis Center, VU University Medical CenterAmsterdam. at the time of the study.
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.
H.V. has received research grants from Pfizer, Novartis and Merck
Serono, and speaker honoraria from Novartis, all paid to his institution.
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.

References

Altmann, D.R., Jasperse, B., Barkhof, F., Beckmann, K., Filippi, M., Kappos, L.D., Molyneux, P., Polman, C.H., Pozzilli, C., Thompson, A.J., Wagner, K., Yousry, T.A., Miller, D.H., 2009. Sample sizes for brain atrophy outcomes in trials for secondary progressive multiple sclerosis. Neurology 72, 595–601.

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

Barkhof, F., Calabresi, P.A., Miller, D.H., Reingold, S.C., 2009. Imaging outcomes for neuroprotection and repair in multiple sclerosis trials. Nature Rev Neurology 5, 256-266.

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.

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.

Clarkson, M.J., Ourselin, S., Nielsen, C., Leung, K.K., Barnes, J., Whitwell, J.L., Gunter, J.L., Hill, D.L., Weiner, M.W., Jack Jr., C.R., Fox, N.C., 2009. Comparison of phantom and registration scaling corrections using the ADNI cohort. Neuroimage 47, 1506–1513.

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.

Cover, KS., van Schijndel, RA., Popescu, V., van Dijk, BW., Redolfi, A., Knol, KL., Frisoni, GB., Barkhof, F., Vrenken H., 2014. The SIENA/FSL whole brain atrophy algorithm is no more reproducible at 3 T than 1.5 T for Alzheimer′s disease. Psychiatry Res. 224, 14–21.

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.

de Jong, L.W., van der Hiele, K., Veer, I.M., Houwing, J.J., Westendorp, R.G.J., Bollen, E.L.E.M., de Bruin, P.W., Middelkoop, H.A.M., van Buchem, M.A., van der Grond, J., 2008. Strongly reduced volumes of putamen and thalamus in Alzheimers disease: an MRI study. Brain 131, 3277-3285.

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.

Evans, M.C., Barnes, J., Nielsen, C., Kim, L.G., Clegg, S.L., Blair, M., Leung, K.K., Douiri, A., Boyes, R.G., Ourselin, S., Fox, N.C., 2010. Volume changes in Alzheimer's disease and mild cognitive impairment: cognitive associations. Euro Radiology 20, 674-682.

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.

Fox, N.C., Freeborough, P.A., 1997. Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease. J. Magn. Reson. Imaging 7, 1069–1075.

Freeborough, P.A., Fox, N.C., 1997. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Trans. Med. Imaging 16, 623–629.

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

Frisoni G.B., Redolfi A., Manset D., Rousseau M.E., Toga A., Evans AC. 2011. Virtual imaging laboratories for marker discovery in neurodegenerative diseases. Nat. Rev. Neurol. 7, 429-438.

Frisoni G.B., Jack C.R., Bocchetta M, Bauer C., Frederiksen K.S., Liu Y, Preboske G., Swihart T., Blair M., Cavedo E., Grothe M.J., Lanfredi M., Martinez O., Nishikawa M., Portegies M., Stoub T., Ward C., Apostolova L.G., Ganzola R., Wolf D., Barkhof F., Bartzokis G., DeCarli C., Csernansky J.G., deToledo-Morrell L., Geerlings M.I., Kaye J., Killiany R.J., Lehéricy S., Matsuda H., O'Brien J., Silbert L.C., Scheltens P., Soininen H., Teipel S., Waldemar G., Fellgiebel A., Barnes J., Firbank M., Gerritsen L., Henneman W., Malykhin N., Pruessner J.C., Wang L., Watson C., Wolf H., deLeon M., Pantel J., Ferrari C., Bosco P., Pasqualetti P., Duchesne S., Duvernoy H., Boccardi M., EADC -European Alzheimer's Disease Consortium and the ADNI (EADC), Alzheimer's Disease Neuroimaging Initiative (ADNI), 2014. The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. Alzheimers Dement. 14, 02468-6.

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.

Iglesiasa, J.E., , Mert R. Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: A survey. Medical Image Analysis 24, 205–219.

Jack, CR., Petersen, RC., OBrien, PC., Tangalos EG., 1992. MR-Based hippocampal volumetery in the diagnosis of Alzheimer’s-disease. Neurology 42, 183-188.

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.

Keihaninejad, S., Heckemann, R.A., Fagiolo, G., Symms, M.R., Hajnal, J.V., Hammers, A., 2010. A robust method to estimate the intracranial volume across MRI field strengths (1.5T and 3T). Neuroimage 50, 1427-1437.

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.

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.

Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, Parikshak N, Toga AW, Jack CR Jr, Schuff N, Weiner MW, Thompson PM; Alzheimer's Disease Neuroimaging Initiative, 2009. Automated mapping of hippocampal atrophy in 1-year repeat MRI data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Neuroimage 45, S3-15.

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.

Mulder, E.R., de Jong, R.A., Knol, D.L., van Schijndel, R.A., Cover, K.S., Visser, P.J., Barkhof, F., Vrenken, H., 2014. Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 92;169-181.

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.

Popescu, V., Battaglini, M., Hoogstrate, W.S., Verfaillie, S.C.J., Sluimer, I.C., van Schijndel, R.A., van Dijk, B.W., Cover, K.S., Knol, D.L., Jenkinson, M., Barkhof, F., de Stefano, N., Vrenken, H., MAGNIMS Study Grp. Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. NeuroImage 2012;61:1484-1494.

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

Rudick, R.A., Fisher, E., Lee, J.C., Simon, J., Jacobs, L., 1999. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. Neurology 53, 1698–1704.

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.

Schneider, J.A., Arvanitakis, Z., Leurgans, S.E., Bennett, D.A. 2009. The Neuropathology of Probable Alzheimer Disease and Mild Cognitive Impairment. Annals of Neurology 66, 200-208.

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.

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, 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.

Sluimer, J.D., van der Flier, W.M., Karas, G.B., van Schijndel, R., Barnes, J., Boyes, R.G., Cover, K.S., Olabarriaga, S.D., Fox, N.C., Scheltens, P., Vrenken, H., Barkhof, F., 2009. Accelerating regional atrophy rates in the progression from normal aging to Alzheimer's disease. European Radiology 19, 2826-2833.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17, 479–489.

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

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.

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.

Watson, P., Head, K., Pitiot, A., Morris, P., Maughan, R.J., 2010. Effect of Exercise and Heat-Induced Hypohydration on Brain Volume. Med. Sci. in Sports and Exercise 42, 2197-2204.

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.

Wilkinson D, Fox NC, Barkhof F, Phul R, Lemming O, Scheltens P. Memantine and brain atrophy in Alzheimer's disease: a 1-year randomized controlled trial. J Alzheimers Dis. 2012;29:459-69.

Wyman BT, Harvey DJ, Crawford K, Bernsteind AM, Carmichael O, Cole PE, Crane PK, DeCarlie C, Fox NC, Gunter JL, Hill D, Killiany RJ, Pachaik C, Schwarzl AJ, Schuff N, Senjem ML, Suhy J, Thompson PM, Weiner M, Clifford R. Jack, Jr., 2013. Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimer’s and Dementia 9, 332–337

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.

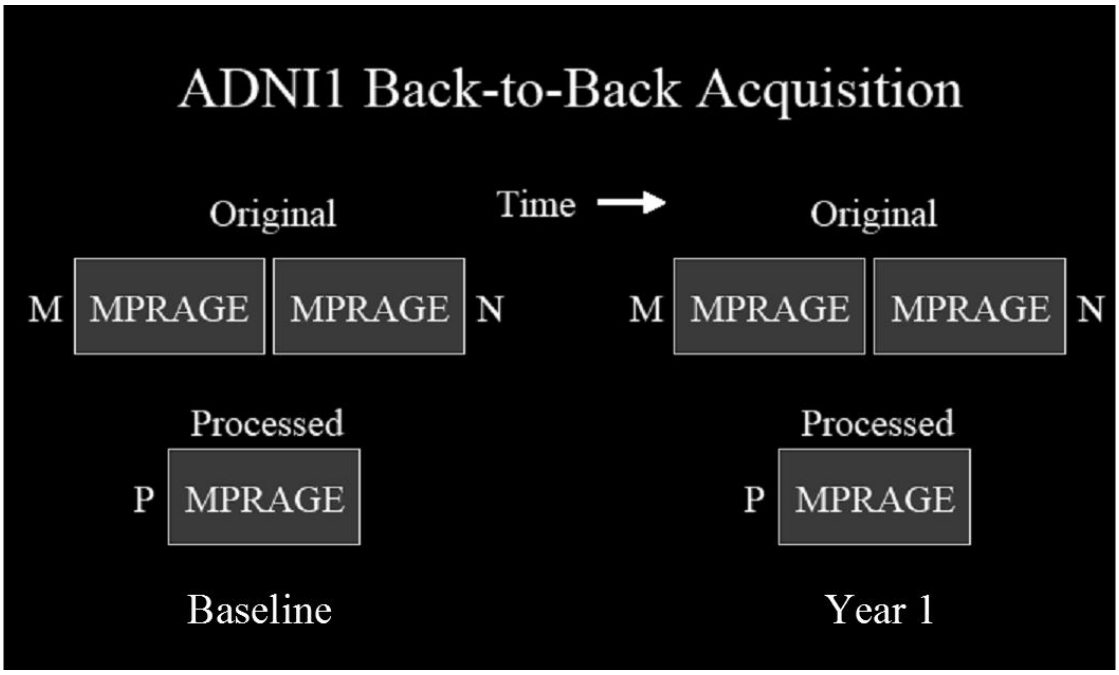

Fig. 1. The 6 MPRAGEs from the ADNI1 study used for each subject in the current study. The 3 MPRAGEs in the left of the figure were generated from the baseline subject visit (A) while the 3 in the right were generated from the one year subject visit (B). The M and N back-to-back (BTB) MPRAGEs are classified as "original" by ADNI as they contain the identical voxel values to those generated by the MRI scanner. The M and N MPRAGEs were acquired BTB with the N acquisition starting within seconds to minutes of the end of the M acquisition. The P MPRAGE is called "processed" by ADNI and was generated by ADNI selecting either M or N for additional processing that is site dependent but includes gradient warp correction. The first acquired MPRAGEs (M) at each of the two subject visits where used for one atrophy rate calculation, the second (N) for a second atrophy rate calculate and the P MPRAGEs for a third.

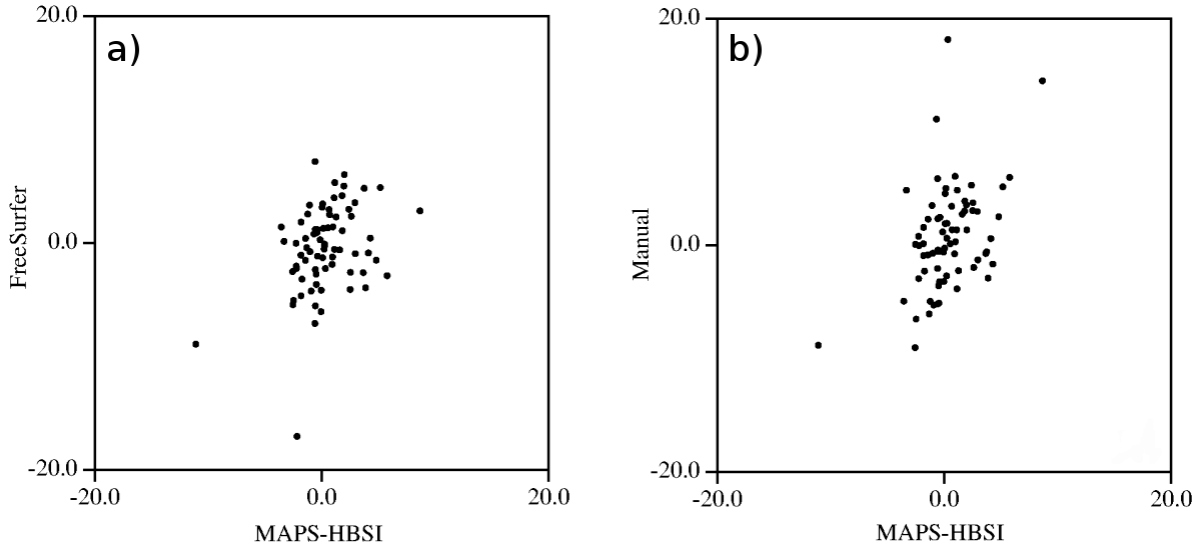

Fig. 2. Scatter plot of the back-to-back (BTB) differences for the manual group of N=75 for the left and right hippocampus of the MAPS-HBSI method versus both the FreeSurfer method in longitudinal mode and the manual segmentation method. As mentioned in the text, a subject's reproducibility is the difference between each subject's two percentage volume changes (PVC). Each of the two PVCs is calculated over one year on each of the pair of MPRAGEs. The smaller spread of the BTB differences for the MAPS-HBSI method (horizontal direction) is clearly evident in each plot as compared to the FreeSurfer and manual methods (vertical directions).

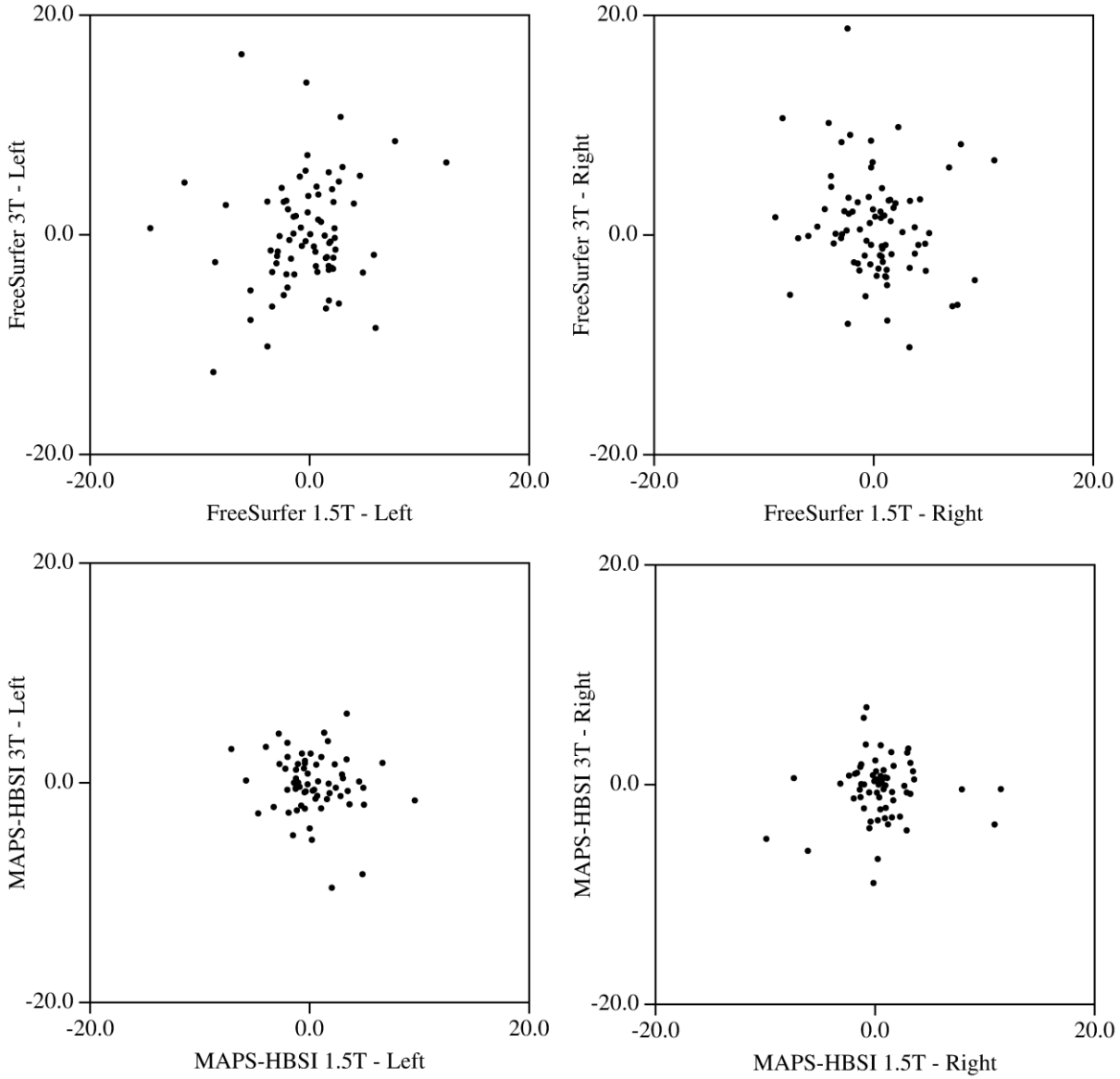

Fig 3. Scatter plots of the BTB differences at 1.5T versus 3T for the hippocampi of longitudinal FreeSurfer and MAPS-HBSI. Each dot represents a subject. The larger scatter of FreeSurfer than MAPS-HBSI - in both the horizontal (1.5T) and vertical (3T) directions - indicates FreeSurfer had poorer reproducibility.

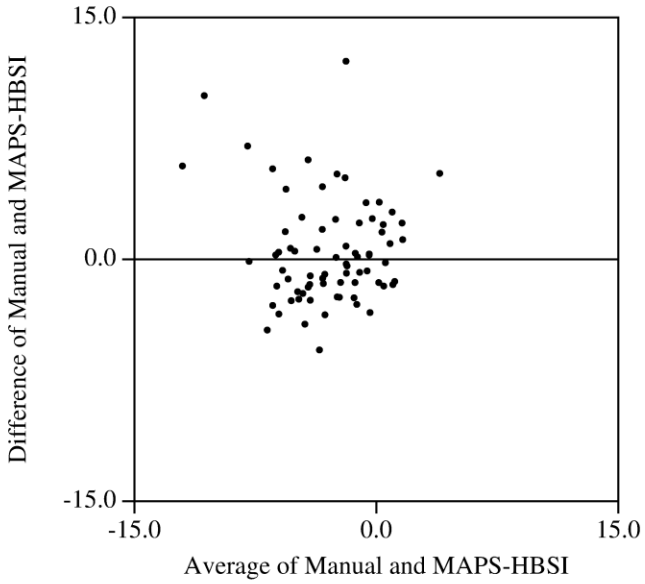

Fig 4. Bland-Altman plot of annualized atrophy rates of manual segmentation versus MAPS-HBSI. The units are percentage points. The clustering of most of the points around the horizontal axis demonstrates the relationship between the atrophy rates of MAPS-HBSI and manual is roughly linear. Thus the manual atrophy rates validate the accuracy of the MAPS-HBSI atrophy rates. The higher the atrophy rate the more negative the value as the annualized percentage volume changes are plotted.

Method	N=562		N=75	
	Left	Right	Left	Right
Manual	N/A	N/A	2.6 (1.0, 4.8)	2.6 (1.0, 4.4)
FSL/FIRST	2.5 (1.1, 4.6)	2.7 (1.4, 5.4)	2.1 (1.1, 4.1)	3.0 (1.5, 5.0)
AdaBoost	2.9 (1.4, 5.0)	3.0 (1.1, 5.5)	2.5 (1.6, 4.6)	4.0 (1.1, 6.1)
FreeSurferC	3.1 (1.5, 5.8)	3.0 (1.2, 5.5)	2.7 (1.4, 4.5)	2.9 (1.3, 5.1)
FreeSurferL	2.1 (1.0, 3.8)	2.0 (0.9, 3.8)	1.9 (0.8, 3.1)	1.7 (1.0, 3.8)
MAPS	2.7 (1.2, 4.8)	2.7 (1.3, 4.8)	2.3 (1.0, 4.7)	2.3 (1.2, 4.2)
MAPS-HBSI	1.3 (0.6, 2.6)	1.3 (0.6, 2.6)	1.3 (0.5, 2.4)	1.1 (0.7, 3.0)

Table 1 Method reproducibility at 1.5T for each method for both left and right hippocampi for both the full N=562 subjects and the manual group of N=75. As mentioned in the text, the method reproducibility is the median of the absolute value of the subjects' BTB differences for the method and a BTB difference is the difference between each subject's BTB percentage volume changes (PVC). The units are percentage points. The interquartile ranges are also displayed. A method with perfect reproducibility would have a value of zero. In all cases the MAPS-HBSI method has the best (smallest) reproducibility.

Method	N=562		N=75	
	Left	Right	Left	Right
Manual	N/A	N/A	0.689 (0.0008)	0.621 (0.0237)
FSL/FIRST	0.685 (<0.00001)	0.712 (<0.00001)	0.685 (0.0011)	0.699 (0.0005)
AdaBoost	0.718 (<0.00001)	0.687 (<0.00001)	0.689 (0.0076)	0.689 (0.0076)
FreeSurferC	0.725 (<0.00001)	0.676 (<0.00001)	0.703 (0.0003)	0.662 (0.0035)
FreeSurferL	0.612 (<0.00001)	0.598 (<0.00001)	0.608 (0.0402)	0.595 (0.0651)
MAPS	0.687 (<0.00001)	0.671 (<0.00001)	0.662 (0.0035)	0.649 (0.0070)

Table 2 Fraction of the subjects for each method at 1.5T whose magnitudes of their BTB differences are bigger than MAPS-HBSI's. The fraction allows the calculation of the statistical significance using the binomial distribution of the difference from MAPS-HBSI of the reproducibility of each method. Each fraction is followed by its p-value in brackets.

Method	HC		MCI		AD	
	L	R	L	R	L	R
FSL/FIRST P	-0.9 (1.0, -2.9)	-1.0 (1.1, -3.4)	-2.8 (-0.2, -5.1)	-2.8 (0.1, -5.1)	-4.0 (-1.0, -5.7)	-2.9 (-0.2, -5.8)
AdaBoost P	-1.0 (0.8, -2.7)	-0.5 (1.5, -2.8)	-3.0 (-0.7, -5.8)	-3.1 (-0.5, -6.0)	-4.3 (1.9, -8.0)	-4.5 (-1.8, -7.3)
FreeSurferC P	-2.7 (0.3, -5.9)	-2.5 (0.0, -5.5)	-3.2 (0.4, -6.6)	-3.2 (-0.2, -6.2)	-4.6 (-0.8, -6.7)	-4.1 (-1.4, -7.0)
FreeSurferL M	-1.2 (0.0, -2.5)	-1.2 (0.0, -2.5)	-2.3 (-0.5, -4.6)	-2.4 (-0.5, -4.7)	-3.9 (-1.6, -6.6)	-3.3 (-1.6, -6.1)
FreeSurferL N	-1.3 (0.0, -2.4)	-1.3 (0.0, -2.4)	-2.5 (-0.3, -4.9)	-2.4 (-0.7, -4.4)	-3.8 (-1.6, -6.6)	-4.1 (-1.6, -6.3)
FreeSurferL P	-1.3 (0.0, -2.7)	-1.0 (0.2, -2.1)	-2.0 (-0.1, -4.5)	-2.3 (-0.6, -4.7)	-3.6 (-1.0, -6.5)	-4.0 (-1.2, -6.4)
MAPS P	-1.2 (1.0, -4.1)	-1.0 (1.2, -2.9)	-3.2 (-1.0, -5.8)	-3.1 (-0.8, -5.4)	-4.8 (-2.3, -7.8)	-3.7 (-2.2, -7.0)
MAPS-HBSI M	-1.4 (-0.2, -2.5)	-1.3 (-0.1, -2.6)	-3.2 (-1.5, -5.5)	-3.0 (-1.0, -5.6)	-5.1 (-2.8, -7.4)	-5.1 (-2.9, -8.1)
MAPS-HBSI N	-1.0 (0.2, -2.3)	-0.9 (0.2, -2.2)	-3.0 (-0.8, -5.2)	-3.0 (-1.0, -5.5)	-5.2 (-2.6, -7.9)	-4.4 (-2.1, -7.3)
MAPS-HBSI P	-1.1 (0.2, -2.5)	-0.9 (0.1, -2.2)	-3.0 (-1.0, -5.1)	-3.3 (-1.2, -5.3)	-5.1 (-2.5, -7.9)	-5.0 (-2.3, -7.6)

Table 3 The median of the atrophy rates at 1.5T for the HC, MCI and AD subsets of N=562 for the methods along with their interquartile ranges. The values are annualized percentage change and are in units of percentage points. The atrophy rates for the cross sectional methods are only given for P to save space.

Method	1.5T		3T	
	Left	Right	Left	Right
FSL/FIRST	2.5 (1.0, 4.5)	3.2 (1.3, 5.3)	2.5 (1.0, 4.5)	3.2 (1.3, 5.3)
FreeSurferC	3.7 (1.1, 6.6)	2.5 (1.1, 5.5)	3.8 (1.6, 6.3)	3.7 (1.5, 6.4)
FreeSurferL	2.1 (1.4, 3.5)	2.0 (0.9, 3.6)	3.0 (1.4, 5.2)	2.5 (1.6, 4.3)
MAPS	2.0 (0.9, 4.4)	2.5 (1.5, 4.2)	3.0 (1.1, 6.4)	2.8 (1.4, 4.9)
MAPS-HBSI	1.4 (0.7, 2.8)	1.1 (0.5, 2.1)	1.6 (0.6, 2.6)	1.5 (0.7, 3.0)

Table 4 Method reproducibility for the 3T group of N=111 similar to Table 1.

Method	Number of Failures N=562 1.5T			Number of Failures N=111 3T	
	M	N	P	M	N
FSL/FIRST	1	4	2	1	0
AdaBoost	0	3	2	0	1
FreeSurferC	1	1	13	1	0
FreeSurferL	1	1	22	1	0
MAPS	7	4	25	0	1
MAPS-HBSI	7	4	25	0	1

Table 5 The number of subjects for each method that failed to yield all the values. Smaller is better.

Optional e-only Supplementary Files

Click here to download Optional e-only Supplementary Files: 20151208N4UBTBBSIManuscript_21_Submitted_TrackChanges.do