Brain Imaging and Behavior (2010) 4:86–95
DOI 10.1007/s11682-010-9088-x

Comparison of Manual and Automated Determination of Hippocampal Volumes in MCI and Early AD

Li Shen · Andrew J. Saykin · Sungeun Kim · Hiram A. Firpi · John D. West · Shannon L. Risacher · Brenna C. McDonald · Tara L. McHugh · Heather A. Wishart · Laura A. Flashman

Received: 18 December 2008 / Accepted: 12 January 2010 / Published online: 13 February 2010

Abstract MRI-based hippocampal volume analysis has been extensively employed given its potential as a biomarker for brain disorders such as Alzheimer's disease (AD), and accurate and efficient determination of hippocampal volumes from brain images is still a challenging issue. We compared an automated method, FreeSurfer (V4), with a published manual protocol for the determination of hippocampal volumes from T1-weighted MRI scans. Our study included MRI data from 125 older adult subjects: healthy controls with no significant cognitive complaints or deficits (HC, n=38), euthymic individuals with cognitive complaints (CC, n=39) but intact neuropsychological performance, and patients with amnestic mild cognitive impairment (MCI, n=37) or a clinical diagnosis of probable AD (AD, n=11). Pearson correlations and intraclass correlation coefficients (ICCs) were calculated to evaluate the relationship between results of the manual tracing and FreeSurfer methods and to estimate their agreement. Results indicated that these two methods derived highly correlated results with strong agreement. After controlling for the age, sex and intracranial volume in statistical group analysis, both the manual tracing and FreeSurfer methods yield similar patterns: both the MCI group and the AD group showed hippocampal volume reduction compared to both the HC group and the CC group, and the HC and CC groups did not differ. These comparisons suggest that FreeSurfer has the potential to be used in automated determination of hippocampal volumes for large-scale MCI/AD-related MRI studies, where manual methods are inefficient or not feasible.

Keywords Magnetic resonance imaging · Segmentation · Hippocampus · Cognitive complaints · Mild cognitive impairment · Alzheimer's disease

L. Shen · A. J. Saykin · S. Kim · H. A. Firpi · J. D. West · S. L. Risacher · B. C. McDonald
IU Center for Neuroimaging, Division of Imaging Sciences, Department of Radiology,
Indiana University School of Medicine,
950 W Walnut St., R2 E124,
Indianapolis, IN 46202, USA

L. Shen (✉)
Center for Computational Biology and Bioinformatics,
Indiana University School of Medicine,
410 West 10th St., Suite 5000,
Indianapolis, IN 46202, USA
e-mail: shenli@iupui.edu

A. J. Saykin · T. L. McHugh · H. A. Wishart · L. A. Flashman
Department of Psychiatry, Dartmouth Medical School, DHMC,
One Medical Center Dr.,
Lebanon, NH 03756, USA

Introduction

The hippocampus has been extensively studied with neuroimaging techniques given its importance in learning and memory and its potential as an imaging biomarker for brain disorders such as Alzheimer's disease (Apostolova et al. 2006; Apostolova et al. 2006; Becker et al. 2006; Csernansky et al. 2005; Saykin et al. 2006; Thompson et al. 2004; Wang et al. 2007; Wang et al. 2006a), epilepsy (Hogan et al. 2006) and schizophrenia (Csernansky et al. 2002; Gerig et al. 2001; Shenton et al. 2002; Wang et al. 2001). While automated and semi-automated methods for segmentation and quantification of the size and shape of the hippocampus were available and used in certain groups (Csernansky et al. 2002; Csernansky et al. 2005; Hogan et al. 2006; Shen et al. 2002; Wang et al. 2007; Wang et al. 2006b), in most studies (Apostolova et al. 2006; Apostolova

et al. 2006; Becker et al. 2006; Bouix et al. 2005; Gerig et al. 2001; McHugh et al. 2007; Saykin et al. 2006; Shenton et al. 2002; Thompson et al. 2004; Yushkevich et al. 2007), researchers employed manual tracing of hippocampal boundaries from magnetic resonance imaging (MRI) scans using various software tools (e.g., Analyze (Mayo Clinic 2008), BRAINS (Iowa Mental Health Clinical Research Center 2008), 3D Slicer (NAMIC 2008), ITK-SNAP (Yushkevich et al. 2006)).

Accurate and efficient MRI segmentation of the hippocampus is still a challenging issue. Although experienced anatomic tracers can be reliable, manual segmentation is a time consuming process and may not be feasible for large-scale neuroimaging studies. For example, the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al. 2005) is a landmark investigation sponsored by the NIH and industrial partners designed to collect longitudinal neuroimaging, biological and clinical information from 800 participants that will track the neural correlates of memory loss from an early stage. All ADNI participants have clinical and cognitive assessments and 1.5 Tesla structural MRI every 6 or 12 months for 2–3 years. Manual segmentation is therefore not ideal given the thousands of MRI scans to be analyzed.

A feasible strategy for hippocampal segmentation in large-scale studies should minimize human interaction involved in the processing pipeline. Diffeomorphic mapping is a notable method for semi-automatic segmentation that has been used by Csernansky and colleagues in many hippocampal studies (Csernansky et al. 2002; Csernansky et al. 2005; Hogan et al. 2006; Wang et al. 2007; Wang, et al, 2006a). FreeSurfer (Dale et al. 1999; Fischl et al. 2002; Fischl et al. 1999) (available at http://surfer.nmr.mgh.harvard.edu/) is an automated software tool for whole brain segmentation and cortical parcellation. Because FreeSurfer is freely available on the web, it has been widely used in the neuroimaging field. One recent study (Tae et al. 2008) reported a comparison of hippocampal volumes measured using the authors' manual method and FreeSurfer V3.04 on 21 female patients with major depressive disorder (MDD) and 20 controls. The intraclass correlation coefficients (ICCs, calculated based on Cronbach's alpha) for testing inter-rater reliability were at a level of 0.85 for both left and right hippocampi (no p values were reported). In this work, we examine data from a different and larger cohort using a newer version of FreeSurfer, and perform a similar analysis comparing hippocampal volume determined by FreeSurfer V4 and a published manual protocol (McHugh et al. 2007).

Materials and methods

In this report, we analyzed data from a well characterized cohort recruited from a study of memory circuitry in mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Here we briefly describe this data set, the two hippocampal segmentation methods examined in the paper, and our methods for comparison.

Participants

Baseline hippocampal data were available for 125 subjects from four categories: healthy older adults with no significant cognitive complaints or deficits (HC, n=38), euthymic older adults with cognitive complaints (CC, n=39) but intact neuropsychological performance, older adults with amnestic MCI (MCI, n=37), and adults with a clinical diagnosis of probable AD (AD, n=11). Demographic characteristics are shown in Table 1.

Structural MRI volume scans were obtained on a GE Signa 1.5 T Horizon LX magnet with echo speed gradients using a standard head RF coil. A T1-weighted three-dimensional spoiled gradient echo (SPGR) coronal volume was acquired. Parameters were TR=25, TE=3 or min, flip angle=40 degrees, 1 NEX, and slice thickness=1.5 mm (no skip), yielding 124 contiguous slices with a 24-cm field of view and a 256 × 256 matrix with 0.9375 mm in-plane resolution. Further details about this data set are available in (Saykin et al. 2006) and in (Wang et al. 2006b).

Manual and automated segmentation

Hippocampal and intracranial boundaries were obtained using (1) a manual protocol reported in (McHugh et al. 2007) using the BRAINS software package (Iowa Mental

Table 1 Sample Characteristics and ANOVA results for testing between group differences

	Age (mean±std, in years)	Education (mean±std, in years)	Sex (M, F)
Healthy Controls (HC, n=38)	70.6±5.2	16.7±2.7	11, 27
Cognitive Complaint (CC, n=39)	72.8±6.1	16.5±2.6	15, 24
Mild Cognitive Impairments (MCI, n=37)	72.7±7.1	16.3±3.2	22, 15
Alzheimer's Disease (AD, n=11)	75.6±6.8	14.4±3.5	5, 6
ALL (n=125)	72.3±6.3	16.3±2.9	53, 72
df, F, p (between group difference)	3, 2.149, 0.098	3, 1.955, 0.124	3, 2.580, 0.057

Health Clinical Research Center 2008), and (2) a fully automated method using the FreeSurfer V4 package (Dale et al. 1999; Fischl et al. 2002; Fischl et al. 1999).

In manual segmentation (Fig. 1), images were reformatted into isotropic 1-mm voxels and resampled into the plane perpendicular to the long axis of the hippocampus using the BRAINS software (Iowa Mental Health Clinical Research Center, 2008). Manual traces were performed in the coronal plane with reference to markings placed in the axial and sagittal views to guide boundary determination. This protocol followed the guidelines of (Jack 1994; Watson et al. 1992) in segmenting the hippocampus with refinements in defining the posterior and anteromedial extents (Andreasen et al. 1993; Torres et al. 1997). Key features included: (1) tracing in the plane of the long axis of the hippocampus to obtain a better view of the hippocampal tail; (2) the "telegraphing" function was used to transfer boundary definition obtained in the axial and sagittal views to the coronal view for guiding boundary determination; and (3) key medial temporal lobe structures were traced together to eliminate boundary overlap. Total incracranial volume (ICV) was also traced to include the cortex, cerebellum, brainstem, and cerebrospinal fluid. Complete details about the manual segmentation protocol and boundaries are available in (McHugh et al. 2007). A 3D binary image was reconstructed from each set of 2D hippocampal segmentation results. Two sample manual segmentation results are shown in the left column of Fig. 2.

In automated segmentation, FreeSurfer was employed to automatically label subcortical tissue classes using an atlas-based Bayesian segmentation procedure. A subject-independent probabilistic atlas was pre-computed from a training set of subjects whose brains were manually labeled. These labels together with intensity values were mapped into Talairach space to enable voxelwise correspondence across subjects. A probabilistic atlas was then created by computing three types of probabilities for each voxel in Talairach space to encode probabilities of class labels, functions of neighborhood patterns and profiles of intensity values. Once the atlas was available, an individual brain MRI could be segmented as follows. Preprocessing steps included an affine registration with Talairach space, intensity normalization, skull strip, and a high dimensional nonlinear volumetric alignment to the Talairach atlas. FreeSurfer calculated the probability of a class at each voxel location as the probability that the given class appeared at that location in the training set times the likelihood of getting the subject-specific intensity value from that class. An initial segmentation was generated by assigning each point to the class for which the above probability was greatest. The neighborhood function was then used to recalculate the class probabilities and re-segment the data using the new class probabilities. This procedure was repeated until the result converged. Details for the FreeSurfer method were reported in (Dale et al. 1999; Fischl et al. 2002; Fischl et al. 1999). The complete FreeSurfer analysis pipeline was performed without any manual intervention. The FreeSurfer V4 package

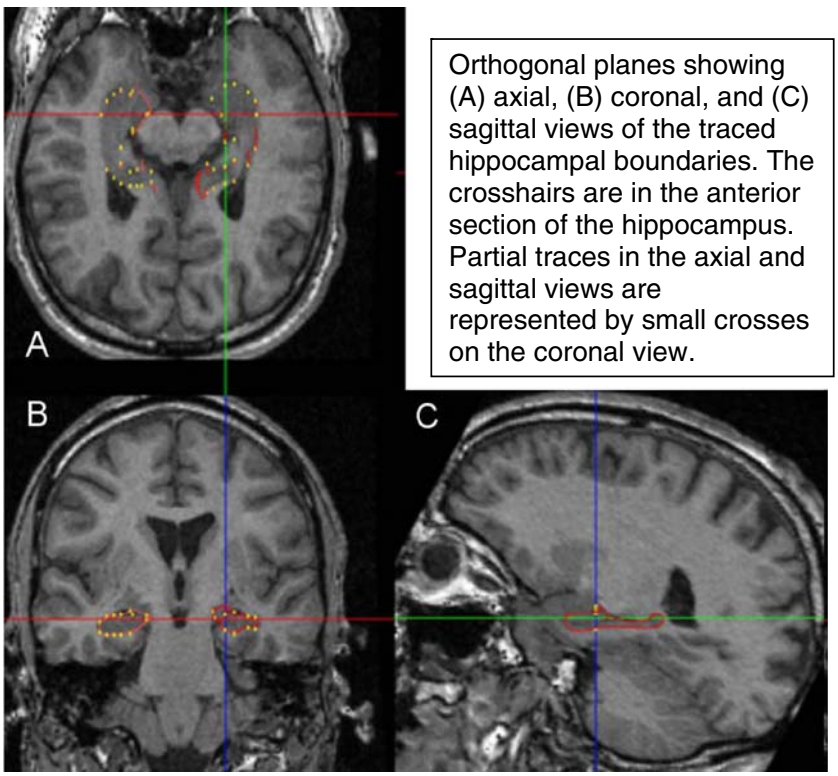

Fig. 1 Manual segmentation using the BRAINS software. (Hippocampal volume and shape analysis in an older adult population by McHugh TL, Saykin AJ, Wishart HA, Flashman LA, Cleavinger HB, Rabin LA, Mamourian AC, and Shen L, The Clinical Neuropsychologist, 21(1):130–145, January 2007, reprinted by permission of the publisher (Taylor & Francis Ltd, http://www.tandf.co.uk/journals))

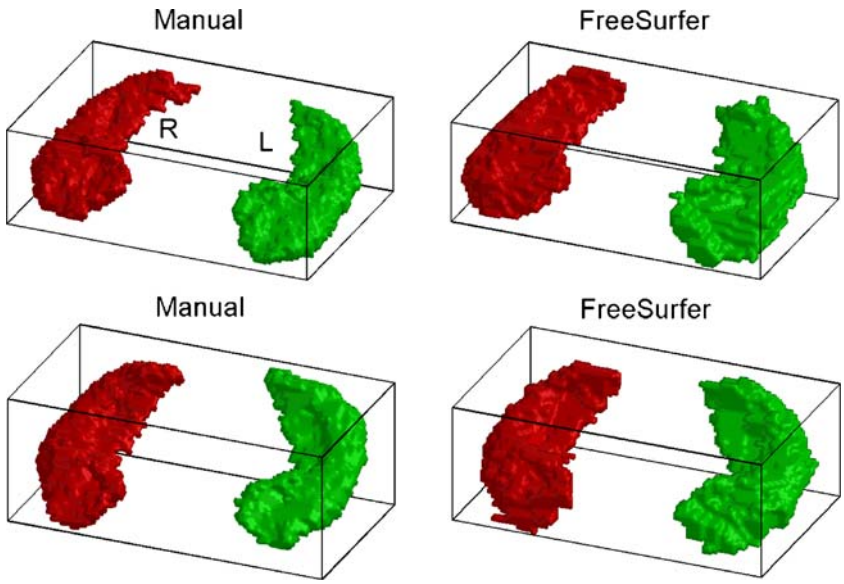

Fig. 2 Sample segmentation results generated by manual tracing (*left*) and FreeSurfer (*right*): Each hippocampus is described by a binary image and the corresponding voxel surface is displayed. Both results have been transformed back to the original space

(v4.0.2) was installed on Quarry (a supercomputer located at Indiana University), an IBM HS21 Bladeserver cluster running Red Hat Linux. The estimated left and right hippocampal volumes were extracted directly from the aseg.stats files. The estimated total intracranial volume (ICV) was re-calculated using the talairach.xfm (new method used in post v4.2.0) instead of the talaiarch_with_skull.lta transform (old method), which was found to be more accurate according to the ICV page on the FreeSurfer website (FreeSurfer Wiki 2009). Further details about the FreeSurfer method for ICV estimation were available in (Buckner et al. 2004).

A sample FreeSurfer volume-based labeling is shown in Fig. 3. Based on each label map, we extracted the left and right hippocampi as two 3D binary images. To visually compare the FreeSurfer and manual tracing results, we transformed both results back to the original image space. Shown in Fig. 2 are two sample manual tracing results (left) and the corresponding FreeSurfer results (right). FreeSurfer results tend to be larger and more inclusive than manual ones and contain several local excursions on the surfaces.

Methods for comparison

Statistical analysis was performed using SPSS 16.0 (SPSS, Chicago, IL). Pearson correlations and interclass correlation coefficients (ICCs) were calculated to evaluate the relationship between volume generated by FreeSurfer and the manual tracing method and to estimate their agreement. Given a variety of available ICC measures that may yield different values for the same data, we briefly describe below (1) the goal of this study and (2) how to choose an appropriate ICC for our reliability test to achieve the goal.

Let D_1 be the data used in this study and D_2 be the data we will collect in the future. For D_1, we have both FreeSurfer and manual tracing results available. Although our manual tracer is experienced and reliable, manual tracing is a very time consuming process and won't be applicable for our future data set D_2 that is much larger than D_1. Therefore, our plan is to use FreeSurfer instead of manual tracing to process D_2. The goal of this study is to perform a feasibility test for our plan and make sure that FreeSurfer can produce accurate results. Since running FreeSurfer on the same data multiple times always generates the same result, the intra-rater reliability for FreeSurfer is 1.0. Therefore, the focus of this study is to

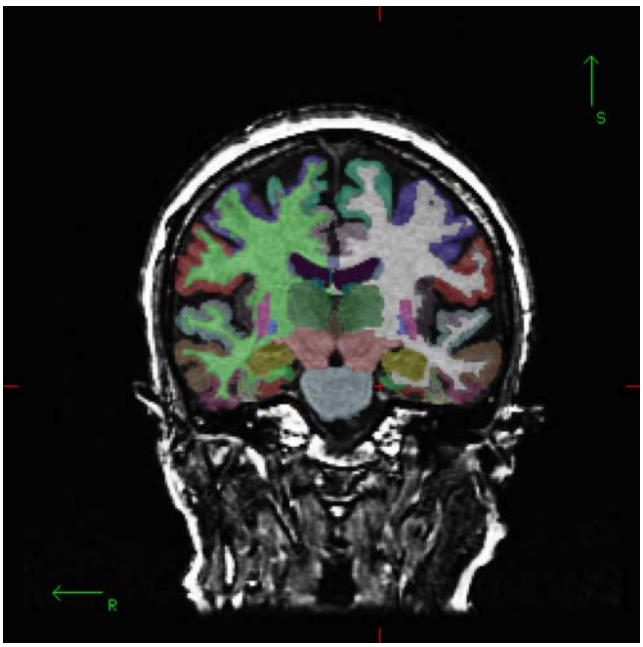

Fig. 3 Automated segmentation using the FreeSurfer software: Left and right hippocampi are shown in yellow

examine the inter-rater reliability by comparing FreeSurfer and manual results using the existing data D_1.

Thus, in our statistical scenario, we aim to collect all the data (i.e., $D_1 \cup D_2$) using one rater (i.e., FreeSurfer) and currently have two raters (i.e., FreeSurfer and manual tracing) on a subset of the data (i.e., D_1) for purposes of estimating inter-rater reliability measured by ICC. We followed the procedure suggested in (McGraw and Wong 1996) to select an appropriate ICC for this type of reliability analysis. We employed a two-way mixed model, since the two raters (i.e., FreeSurfer and manual tracer) were a fixed effect while the target ratings (i.e., hippocampal volumes) were a random effect in our study. We tested the single measure reliability instead of the average measure reliability, because our goal was to evaluate the reliability of the ratings for a typical single rater (i.e., FreeSurfer) rather than the mean of all the ratings. We selected "consistency" as the model type instead of "absolute agreement", since we were more interested in seeing the consistency of the relative standing of the measures than absolute agreement between two raters. In short, the SPSS option for this type of ICC analysis can be summarized as the "two-way mixed model of single measure intraclass correlation with consistency type". The formula for computing this type of ICC is available in (McGraw and Wong 1996) and can be found in the formula entry for Case 3 model ICC(C,1) in Table 4 of (McGraw and Wong 1996). In our experiments, we performed the above analysis within each group as well as across groups.

In addition to comparing the FreeSurfer results directly with the manual tracing results to estimate their agreement, we performed a group analysis (SPSS General Linear Model) separately for each method examining whether hippocampal volumes differ significantly among the HC, CC, MCI, and AD groups. Our goal was to see if the FreeSurfer and manual tracing methods would yield a similar pattern of group differences. In these analyses, we also tested effects of age, education, sex and ICV on volume measures determined by each method, and then involved age, sex and ICV as covariates in our linear models.

Results

We analyzed both raw hippocampal volumes and "relative hippocampal volumes to ICV" that were calculated as the raw volumes divided by the corresponding ICVs. Fig. 4 shows the plots of the FreeSurfer measures against the

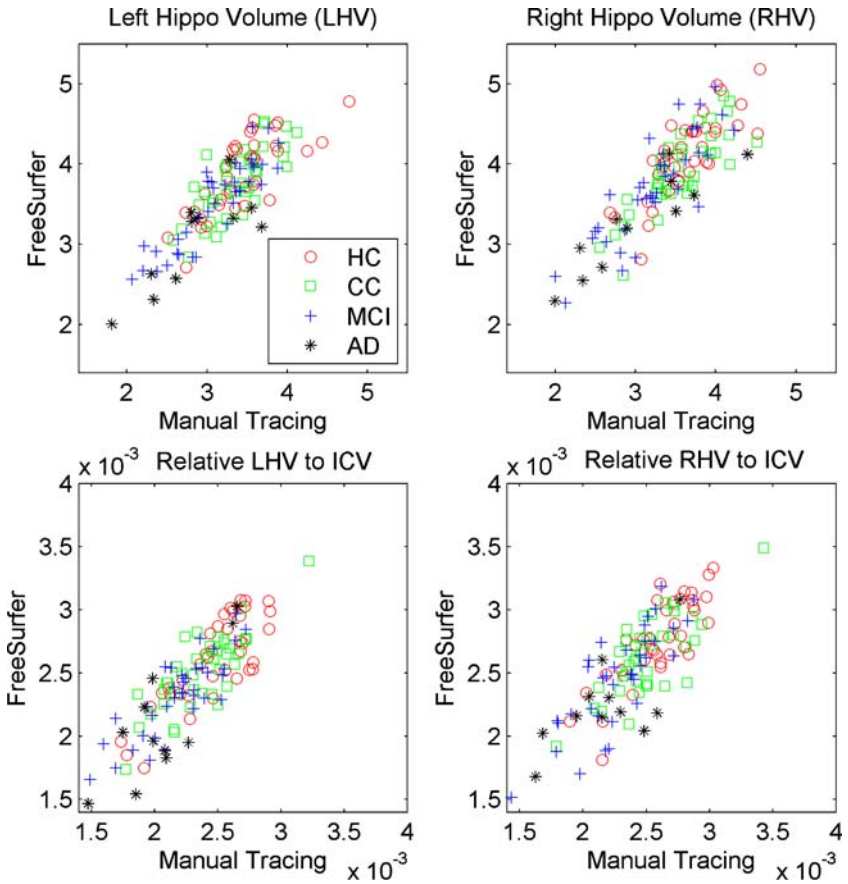

Fig. 4 Visual comparison between hippocampal volumes determined by FreeSurfer and manual tracing methods: Both raw hippocampal volumes (LHV, RHV) and relative hippocampal volumes to ICV (i.e., volume/ICV) were plotted. Pearson correlation analysis results are available in Table 2

Table 2 Pearson correlation coefficients r with *p*-value between hippocampal volumes determined by FreeSurfer and manual tracing methods. Both raw hippocampal volumes (LHV, RHV) and 'relative' hippocampal volumes (i.e., volume divided by the corresponding ICV) were analyzed; see Fig. 4 for scatter plots. Intracranial volumes (ICVs) determined by FreeSurfer and manual tracing methods were also compared

	All (n=125)	HC (n=38)	CC (n=39)	MCI (n=37)	AD (n=11)
LHV	r=0.842	r=0.771	r=0.757	r=0.895	r=0.808
	p<0.001	p<0.001	p<0.001	p<0.001	p=0.003
RHV	r=0.850	r=0.803	r=0.826	r=0.849	r=0.901
	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001
Relative LHV	r=0.850	r=0.817	r=0.797	r=0.862	r=0.813
	p<0.001	p<0.001	p<0.001	p<0.001	p=0.002
Relative RHV	r=0.832	r=0.819	r=0.796	r=0.834	r=0.660
	p<0.001	p<0.001	p<0.001	p<0.001	p=0.027
ICV	r=0.929	r=0.919	r=0.930	r=0.942	r=0.931
	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001

manual tracing measures. A visual inspection of Fig. 4 indicates that FreeSurfer volumes as unadjusted numbers are systematically larger than manual results. After adjusting ICV, this trend gets substantially reduced. We computed and examined Pearson correlations across groups and within each group for not only hippocampal volumes but also ICV measures, and Table 2 summarizes the results. The correlations within each group and across groups for absolute and relative hippocampal measures are all significant at the level of $r \geq 0.757$ except r=0.66 for relative RHV in the AD group. The correlations for ICV are all significant at a higher level of $r \geq 0.919$.

We also computed and examined ICCs for testing inter-rater reliability across groups and within each group. Table 3 provides inter-rater reliability results for the single measure ICC for the entire set and each group respectively. As mentioned earlier, since our goal is to test if FreeSurfer is a valid alternative method, single measure ICCs under a two-way mixed model with consistency type are most appropriate. These results show reliabilities are significant

Table 3 Intraclass Correlation Coefficients (ICCs) within group and between groups: Two-way mixed effects model where sample (person) effects are random and measure (manual, automated) effects are fixed. A "relative volume to ICV" is calculated as the original volume divided by the corresponding ICV

		Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
			Lower Bound	Upper Bound	Value	df1	df2	Sig
All (n=125)	Left Hippo Vol (LHV)	0.838	0.777	0.883	11.349	124	124	<0.001
	Right Hippo Vol (RHV)	0.845	0.786	0.888	11.865	124	124	<0.001
	Relative LHV to ICV	0.845	0.786	0.889	11.907	124	124	<0.001
	Relative RHV to ICV	0.825	0.760	0.874	10.437	124	124	<0.001
HC (n=38)	Left Hippo Vol (LHV)	0.771	0.601	0.874	7.716	37	37	<0.001
	Right Hippo Vol (RHV)	0.793	0.637	0.887	8.672	37	37	<0.001
	Relative LHV to ICV	0.811	0.666	0.897	9.586	37	37	<0.001
	Relative RHV to ICV	0.805	0.655	0.893	9.240	37	37	<0.001
CC (n=39)	Left Hippo Vol (LHV)	0.752	0.574	0.862	7.054	38	38	<0.001
	Right Hippo Vol (RHV)	0.823	0.687	0.903	10.288	38	38	<0.001
	Relative LHV to ICV	0.796	0.644	0.888	8.811	38	38	<0.001
	Relative RHV to ICV	0.787	0.629	0.882	8.368	38	38	<0.001
MCI (n=37)	Left Hippo Vol (LHV)	0.893	0.802	0.943	17.697	36	36	<0.001
	Right Hippo Vol (RHV)	0.834	0.701	0.911	11.068	36	36	<0.001
	Relative LHV to ICV	0.862	0.748	0.926	13.490	36	36	<0.001
	Relative RHV to ICV	0.823	0.682	0.905	10.290	36	36	<0.001
AD (n=11)	Left Hippo Vol (LHV)	0.807	0.433	0.944	9.387	10	10	0.001
	Right Hippo Vol (RHV)	0.889	0.640	0.969	16.939	10	10	<0.001
	Relative LHV to ICV	0.763	0.333	0.930	7.436	10	10	0.002
	Relative RHV to ICV	0.660	0.135	0.896	4.882	10	10	0.010

Table 4 Linear regression results on all subjects (n=125). Triple values in each entry represent regression coefficient, t-value, and p-value, respectively. Significant results (p<0.05) are shown in **bold**

	Age	Education	Sex	ICV_MT	ICV_FS
Left_MT	**−0.030,−5.026, <0.001**	0.002, 0.104, 0.918	0.005, 0.048, 0.962	**0.001, 3.911, <0.001**	N/A
Right_MT	**−0.026,−3.978, <0.001**	0.004, 0.230, 0.818	0.107, 0.993, 0.323	**0.002, 4.808, <0.001**	N/A
Left_FS	**−0.043,−6.789, <0.001**	−0.005,−0.317, 0.752	−0.023,−0.232, 0.817	N/A	**0.001, 3.466, 0.001**
Right_FS	**−0.048,−6.893, <0.001**	0.006, 0.328, 0.744	0.079, 0.726, 0.469	N/A	**0.001, 3.861, <0.001**

at levels ranging from 0.75 to 0.89 within group and across groups. The only outlier is the "relative right hippocampal volume (RHV) to ICV" in the AD group (ICC=0.66), whose low ICC level could be caused by a combination of several factors. First, the sample size (n=11) is relatively small compared with other groups. Second, the quality of ICV measures could be an additional reason, given that we had ICC=0.889 for RHV without adjusting for ICV. For each participant, FreeSurfer and manual methods determine their own ICV values that are subsequently used for calculating their corresponding relative RHVs. Although the correlation between FreeSurfer and manual ICVs is significant at a level of r=0.929 (see Table 2), small disturbance on either FreeSurfer or manual ICV estimate could cause a big change between the ICC for RHV and the ICC for relative RHV to ICV, given such a small sample size (n=11). A similar pattern was also observed in Pearson correlation results (Table 2): r=0.901 for RHV and r=0.660 for relative RHV for the AD group.

In addition to comparing the agreement of FreeSurfer directly with manual tracing, we performed group analysis (SPSS General Linear Model) on data sets measured by both methods to determine if a similar pattern of group differences would be obtained. We first examined the effects of age, education, sex and ICV on the left and right hippocampal volumes determined by both FreeSurfer and manual methods, using a linear regression model; and the results were summarized in Table 4. The following patterns were obtained in both methods: the effects of age and ICV were significant ($p \leq 0.05$) and the effects of education and sex were not significant. After that, we ran our linear models to examine the diagnostic group effect (HC, CC, MCI, and AD) on the hippocampal volume with age, sex and ICV as covariates. Age-, sex- and ICV-adjusted hippocampal volumes were plotted in Fig. 5. Similar between-group differences were detected using manual tracing and FreeSurfer: (1) for manual tracing data, we had degrees-of-freedom (df)=6, F=13.836, p<0.001 for left hippocampus and df=6, F=13.120, p<0.001 for right hippocampus; and (2) for FreeSurfer data, we had df=6, F=17.620, p<0.001 for left and df=6, F=15.788, p<0.001 for right.

More detailed results on pairwise group comparisons were reported in Table 5. For both FreeSurfer and manual tracing methods, the following patterns were observed. The AD group showed hippocampal volume reduction compared to both the HC (left p=0.001 and right p=0.002 for manual tracing data, left and right both p<0.001 for FreeSurfer data) and CC (left p=0.004 and right p=0.011 for manual tracing data, left p<0.001 and right p<0.001 for FreeSurfer data) groups. The MCI group also showed hippocampal volume reduction compared to both the HC (left and right both p<0.001 for manual tracing data, left p=0.001 and right p=0.002 for FreeSurfer data) and CC

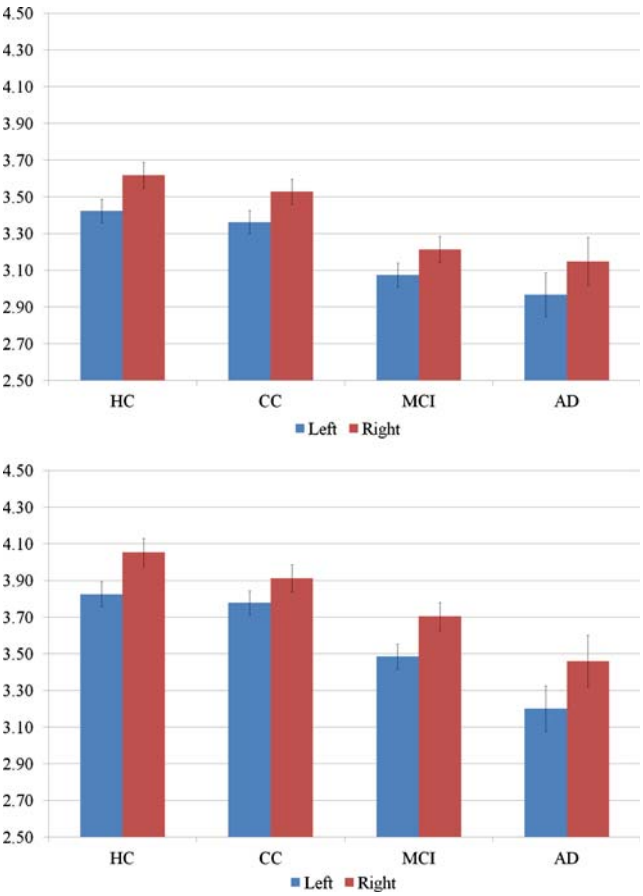

Fig. 5 Estimated marginal means of hippocampal volumes determined by the manual tracing method and the FreeSurfer method. Age, sex, and ICV adjusted means (±Std. Error) for manually traced hippocampal volumes are shown at the top and for FreeSurfer segmented volumes at the bottom

(left and right both p=0.002 for manual tracing data, left p=0.002 and right p=0.053 for FreeSurfer data) groups. The HC and CC groups did not differ (left p=0.495 and right p=0.362 for manual tracing data, left p=0.619 and right p=0.183 for FreeSurfer data).

We did observe a disagreement between FreeSurfer and manual methods for detecting the volume difference of the left hippocampus between MCI and AD: while the difference is significant for FreeSurfer (p=0.044), it is not significant for manual method (p=0.428). Table 5 summarized all the above pairwise comparison results as well as reported the corresponding effect size for each case. Since the sex effect on each hippocampal volume measure was not significant (Table 4), we also tested another linear model without including sex as a covariate. That is, we only controlled the effects of age and ICV in the model, and the results were either identical or extremely close to those reported in Table 5.

We also examined the difference between left and right hippocampal volumes within each group and across groups using paired samples t-tests. Similar left and right differences were detected using manual tracing and FreeSurfer: the left-right hippocampal volume differences were significant within each group and across groups (right>left, $p \leq$ 0.001).

As mentioned earlier, FreeSurfer volumes as unadjusted numbers are systematically larger than manual results (Fig. 4). After adjusting for age, sex and ICV, this systematic difference remains to some extent (Fig. 5). By visual inspection of the entire sample (see Fig. 2 for two examples), we noticed that FreeSurfer results tend to (1) be more inclusive than manual tracing results, especially in the tail region, and (2) have quite a few local excursions on the surfaces. To examine whether this is the cause of the systematic difference shown in Fig. 5 (i.e., FreeSurfer Volume>Manual Volume), detailed analysis of shape differences between these two sets of results would be a useful topic for future investigation.

Discussion

We compared an automated method, FreeSurfer (V4), with a published manual protocol for the determination of hippocampal volume from MRI scans, using data from an existing MCI/AD cohort. Although the two methods might not result in the same volume estimates, our correlation and reliability analyses showed that the hippocampal volumes determined by the manual tracing method and the FreeSurfer method were not only significantly correlated to each

Table 5 Pairwise group comparisons based on estimated marginal means with age, sex, ICV as covariates. Significant results (p<0.05) and large effects (d>0.8) are shown in **bold**

	Diagnosis Comparison	Mean Difference	Std. Error	Sig. (p)	Effect Size (d)[a]
Left_MT	HC-CC	0.061	0.089	0.495	0.156
	HC-MCI	0.349	0.093	**<0.001**	**0.881**
	HC-AD	0.456	0.136	**0.001**	**1.155**
	CC-MCI	0.288	0.091	**0.002**	0.733
	CC-AD	0.395	0.134	**0.004**	**1.008**
	MCI-AD	0.107	0.134	0.428	0.271
Right_MT	HC-CC	0.090	0.098	0.362	0.209
	HC-MCI	0.405	0.102	**<0.001**	**0.936**
	HC-AD	0.469	0.149	**0.002**	**1.088**
	CC-MCI	0.315	0.099	**0.002**	0.735
	CC-AD	0.379	0.146	**0.011**	**0.887**
	MCI-AD	0.064	0.147	0.662	0.149
Left_FS	HC-CC	0.046	0.093	0.619	0.114
	HC-MCI	0.340	0.097	**0.001**	**0.827**
	HC-AD	0.624	0.142	**<0.001**	**1.520**
	CC-MCI	0.294	0.094	**0.002**	0.722
	CC-AD	0.578	0.139	**<0.001**	**1.423**
	MCI-AD	0.284	0.140	**0.044**	0.694
Right_FS	HC-CC	0.141	0.105	0.183	0.305
	HC-MCI	0.350	0.110	**0.002**	0.749
	HC-AD	0.593	0.162	**<0.001**	**1.272**
	CC-MCI	0.209	0.107	0.053	0.452
	CC-AD	0.452	0.158	**0.005**	**0.980**
	MCI-AD	0.243	0.159	0.128	0.523

[a] Cohen's d: "small, d = .2," "medium, d = .5," and "large, d = .8" based on conventional operational definitions in power analysis

other at levels ranging from 0.76 to 0.90 but also showed significant agreement measured by ICCs ranging from 0.75 to 0.89.

To test whether the two methods could detect between group differences with similar sensitivity, we conducted additional statistical analysis to examine the hippocampal volume difference between left and right sides as well as among four groups: HC, CC, MCI and AD. Both the manual tracing method and the FreeSurfer method detected significant differences between left and right hippocampal volumes within group and across groups. Both methods also yielded the following patterns: The AD group and the MCI group showed hippocampal volume reduction compared to both the HC and CC, and the HC and CC groups did not differ. We have hypothesized that the CC group may represent a pre-MCI stage of AD and therefore eventually provide an earlier therapeutic opportunity than MCI as new disease modifying agents are assessed. In this regard it is noteworthy that both manually traced and FreeSurfer segmented hippocampal volumes, adjusted for age, gender and ICV, were significantly reduced relative to HC only in the MCI and AD groups, with the CC group showing an intermediate level between MCI and HC. Although volumetric methods did not identify differences between the HC and CC groups, the CC group showed similar but more significant patterns of decreased gray matter density relative to the HC group on whole brain and hippocampal ROI analyses, and further details are available in Saykin et al (2006).

In summary, the two methods could yield different volume estimates but produce similar capacity for group discrimination, suggesting that the FreeSurfer method may be useful for automated determination of hippocampal volumes for group testing in large-scale MCI/AD-related MRI studies. Reproducibility is also very important in that another use in clinical trials could be as a biomarker for detecting the effects of putative disease modifying treatments.

Although the convergent validity of hippocampal volume determined by FreeSurfer has been tested in this work and other similar studies (e.g., (Tae et al. 2008)), future research should examine alternative situations, such as data from different scanners, of different resolutions, contrast profiles, or with different signal to noise ratios. Another potentially important research direction is to assess if FreeSurfer can also generate more complex morphometric measurements beyond volume (e.g., orientation or shape features) (Shen et al., 2009) or if additional post-processing steps need to be developed in order to achieve such a goal.

Overall, FreeSurfer represents an advance in robust automated volume determination that agrees reasonably well with a manual tracing method for examining the hippocampus. With additional software enhancements, ever improving scan quality, and additional validation, this and related automated methods are poised to make significant contributions to medical image analysis which in turn can provide robust biomarkers for research and clinical applications.

Acknowledgements Supported in part by NIA R01 AG19771, NIBIB R03 EB008674-01, NCI R01 CA101318, U54 EB005149 and NIA P30 AG10133 from the NIH, Foundation for the NIH, and grant #87884 from the Indiana Economic Development Corporation (IEDC). We thank Nick Schmansky and Bruce Fischl of Harvard Medical School and Randy Heiland of Indiana University for help with running FreeSurfer on IU's supercomputers.

References

Andreasen, N. C., Cizadlo, T., Harris, G., Swayze, V., O'Leary, D. S., Cohen, G., et al. (1993). Voxel processing techniques for the antemortem study of neuroanatomy and neuropathology using magnetic resonance imaging. *Journal of Neuropsychiatry, 5*, 121–130.

Apostolova, L. G., Dinov, I. D., Dutton, R. A., Hayashi, K. M., Toga, A. W., Cummings, J. L., et al. (2006). 3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease. *Brain, 129*(Pt 11), 2867–2873.

Apostolova, L. G., Dutton, R. A., Dinov, I. D., Hayashi, K. M., Toga, A. W., Cummings, J. L., et al. (2006). Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. *Archives of Neurology, 63*(5), 693–699.

Becker, J. T., Davis, S. W., Hayashi, K. M., Meltzer, C. C., Toga, A. W., Lopez, O. L., et al. (2006). Three-dimensional patterns of hippocampal atrophy in mild cognitive impairment. *Archives of Neurology, 63*(1), 97–101.

Bouix, S., Pruessner, J. C., Louis Collins, D., & Siddiqi, K. (2005). Hippocampal shape analysis using medial surfaces. *Neuroimage, 25*(4), 1077–1089.

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., et al. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. *NeuroImage, 23*, 724–738.

Csernansky, J. G., Wang, L., Jones, D., Rastogi-Cruz, D., Posener, J. A., Heydebrand, G., et al. (2002). Hippocampal deformities in schizophrenia characterized by high dimensional brain mapping. *American Journal of Psychiatry, 159*(12), 2000–2006.

Csernansky, J. G., Wang, L., Swank, J., Miller, J. P., Gado, M., McKeel, D., et al. (2005). Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. *Neuroimage, 25*(3), 783–792.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis I. Segmentation and surface reconstruction. *Neuroimage, 9*(2), 179–194.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron, 33*(3), 341–355.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage, 9*(2), 195–207.

FreeSurfer Wiki (2009). eTIV—estimated Total Intracranial Volume, aka ICV. *Available at* http://surfer.nmr.mgh.harvard.edu/fswiki/eTIV

Gerig, G., Styner, M., Shenton, M. E., & Lieberman, J. A. (2001, Oct. 14–17). *Shape versus size: Improved understanding of the morphology of brain structures.* Paper presented at the 4th

International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001), Ultrecht, the Netherlands.

Hogan, R. E., Wang, L., Bertrand, M. E., Willmore, L. J., Bucholz, R. D., Nassif, A. S., et al. (2006). Predictive value of hippocampal MR imaging-based high-dimensional mapping in mesial temporal epilepsy: preliminary findings. *AJNR. American Journal of Neuroradiology, 27*(10), 2149–2154.

Iowa Mental Health Clinical Research Center (2008). BRAINS Software Package. *Available at* http://www.psychiatry.uiowa.edu/mhcrc/IPLpages/BRAINS.htm

Jack, C. R. J. (1994). MRI-based hippocampal volume measurements in epilepsy. *Epilepsia, 35*(Suppl 6), S21–S29.

Mayo Clinic (2008). Analyze. *Available at* http://www.mayo.edu/bir/Software/Analyze/Analyze.html

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

McHugh, T. L., Saykin, A. J., Wishart, H. A., Flashman, L. A., Cleavinger, H. B., Rabin, L. A., et al. (2007). Hippocampal volume and shape analysis in an older adult population. *Clin Neuropsychol, 21*(1), 130–145.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America, 15*(4), 869–877. xi–xii.

NAMIC (2008). 3D Slicer Web Page. *Available at* http://www.slicer.org

Saykin, A. J., Wishart, H. A., Rabin, L. A., Santulli, R. B., Flashman, L. A., West, J. D., et al. (2006). Older adults with cognitive complaints show brain atrophy similar to that of amnestic MCI. *Neurology, 67*, 834–842.

Shen, D., Moffat, S., Resnick, S. M., & Davatzikos, C. (2002). Measuring size and shape of the hippocampus in MR images using a deformable shape model. *Neuroimage, 15*(2), 422–434.

Shen, L., Firpi, H. A., Saykin, A. J., & West, J. D. (2009). Parametric surface modeling and registration for comparison of manual and automated segmentation of the hippocampus. *Hippocampus, 19*(6), 588–595.

Shenton, M. E., Gerig, G., McCarley, R. W., Szekely, G., & Kikinis, R. (2002). Amygdala-hippocampal shape differences in schizophrenia: the application of 3D shape models to volumetric MR data. *Psychiatry Research, 115*(1–2), 15–35.

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.*

Thompson, P. M., Hayashi, K. M., De Zubicaray, G. I., Janke, A. L., Rose, S. E., Semple, J., et al. (2004). Mapping hippocampal and ventricular change in Alzheimer disease. *Neuroimage, 22*(4), 1754–1766.

Torres, I. J., Flashman, L. A., O'Leary, D. S., Swayze, V. I., & Andreasen, N. C. (1997). Lack of an association between delayed memory and hippocampal and temporal lobe size in patients with schizophrenia and healthy controls. *Biological Psychiatry, 42*(12), 1087–1096.

Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J. C., et al. (2007). Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. *IEEE Trans Med Imaging, 26*(4), 462–470.

Wang, L., Joshi, S. C., Miller, M. I., & Csernansky, J. G. (2001). Statistical analysis of hippocampal asymmetry in schizophrenia. *Neuroimage, 14*(3), 531–545.

Wang, L., Miller, J. P., Gado, M. H., McKeel, D. W., Rothermich, M., Miller, M. I., et al. (2006a). Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. *Neuroimage, 30*(1), 52–60.

Wang, P., Saykin, A., Flashman, L., Wishart, H., Rabin, L., Santulli, R., et al. (2006b). Regionally specific atrophy of the corpus callosum in AD, MCI and cognitive complaints. *Neurobiology of Aging, 27*(11), 1613–1617.

Watson, C., Andermann, F., Gloor, P., Jones-Gotman, M., Peters, T., Evans, A., et al. (1992). Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. *Neurology, 42*(9), 1743–1750.

Yushkevich, P. A., Detre, J. A., Mechanic-Hamilton, D., Fernandez-Seara, M. A., Tang, K. Z., Hoang, A., et al. (2007). Hippocampus-specific fMRI group activation analysis using the continuous medial representation. *Neuroimage, 35*(4), 1516–1530.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage, 31*(3), 1116–1128.