Comparing Manual and Automatic Segmentation of Hippocampal Volumes: Reliability and Validity Issues in Younger and Older Brains

Elisabeth Wenger,[1]* Johan Mårtensson,[1,2] Hannes Noack,[1,3] Nils Christian Bodammer,[1] Simone Kühn,[1] Sabine Schaefer,[1] Hans-Jochen Heinze,[4,5] Emrah Düzel,[3,4,6] Lars Bäckman,[7] Ulman Lindenberger,[1] and Martin Lövdén[1,7]*

[1]Center for Lifespan Psychology, Max Planck Institute for Human Development, Germany
[2]Department of Psychology, Lund University, Sweden
[3]Institute of Medical Psychology and Behavioral Neurobiology, Tübingen University, Germany
[4]Department of Neurology, Otto-von-Guericke University of Magdeburg, Germany
[5]German Centre for Neurodegenerative Diseases (DZNE), Magdeburg, Germany
[6]Institute of Cognitive Neuroscience, University College London, United Kingdom
[7]Aging Research Center, Karolinska Institutet and Stockholm University, Sweden

Abstract: We compared hippocampal volume measures obtained by manual tracing to automatic segmentation with FreeSurfer in 44 younger (20–30 years) and 47 older (60–70 years) adults, each measured with magnetic resonance imaging (MRI) over three successive time points, separated by four months. Retest correlations over time were very high for both manual and FreeSurfer segmentations. With FreeSurfer, correlations over time were significantly lower in the older than in the younger age group, which was not the case with manual segmentation. Pearson correlations between manual and FreeSurfer estimates were sufficiently high, numerically even higher in the younger group, whereas intra-class correlation coefficient (ICC) estimates were lower in the younger than in the older group. FreeSurfer yielded higher volume estimates than manual segmentation, particularly in the younger age group. Importantly, FreeSurfer consistently overestimated hippocampal volumes independently of manually assessed volume in the younger age group, but overestimated larger volumes in the older age group to a less extent, introducing a systematic age bias in the data. Age differences in hippocampal volumes were significant with FreeSurfer, but not with manual tracing. Manual tracing resulted in a significant difference between left and right hippocampus (right > left), whereas this asymmetry

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: Max Planck Society; Sofja Kovalevskaja Award, the Alexander von Humboldt foundation, the German Federal Ministry for Education and Research (BMBF); German Research Council; BMBF; Swedish Research Council; Swedish Brain Power; Alexander von Humboldt Research Award; af Jochnick Foundation; International Max Planck Research School "The Life Course: Evolutionary and Ontogenetic Dynamics" (LIFE).

*Correspondence to: Elisabeth Wenger, Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: Wenger@mpib-berlin.mpg.de (or) Martin Lövdén, Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: Martin.lovden@ki.se

Disclosure Statement: No conflicts of interest.

Received for publication 3 December 2013; Accepted 14 January 2014.

DOI 10.1002/hbm.22473

Published online 14 February 2014 in Wiley Online Library (wileyonlinelibrary.com).

effect was considerably smaller with FreeSurfer estimates. We conclude that FreeSurfer constitutes a feasible method to assess differences in hippocampal volume in young adults. FreeSurfer estimates in older age groups should, however, be interpreted with care until the automatic segmentation pipeline has been further optimized to increase validity and reliability in this age group. *Hum Brain Mapp 35:4236–4248, 2014.* © 2014 Wiley Periodicals, Inc.

Key words: hippocampus; FreeSurfer; manual segmentation; left right asymmetry; aging

INTRODUCTION

Given its importance in memory and learning, the hippocampus (Hc) has been extensively studied with neuroimaging techniques in the last decade [e.g., Squire et al., 2004]. Measurement of hippocampal volume has become an important diagnostic tool in clinical settings [e.g., Müller et al., 2010], and volume differences have been revealed for a variety of neurological and psychiatric disorders [Bremner et al., 1995; Chetelat and Baron, 2003; Heckers, 2001; Jack et al., 1999; Sheline et al., 1996]. Measuring hippocampal volume is also important in studies of healthy adults [Bhatia et al., 1993; Honeycutt and Smith, 1995]. The Hc is considered one of the very few brain regions that preserve their potential for neurogenesis into adulthood and aging [Christie and Cameron, 2006]. It is therefore a valuable model for studying both the potential for neuroplasticity as well as its limits in mature brains. Animal studies have demonstrated hippocampal changes in response to experience [Churchhill et al., 2002; Jessberger and Gage, 2008; Kempermann et al., 2002; Kronenberg et al., 2006; Rosenzweig and Bennett, 1996; van Praag et al., 2000]. In the last years, neuroplasticity within the hippocampal formation has also been shown in humans. In their seminal studies, Maguire et al. showed that London taxi drivers have a larger posterior hippocampus region compared to controls and that successful spatial knowledge acquisition is related to hippocampal growth [Maguire et al., 2000, 2006; Woollett and Maguire, 2011]. Other types of learning can trigger changes in hippocampal volume as well, for example learning how to juggle over a period of 3 months [Boyke et al., 2008], studying for a final medical exam [Draganski et al., 2006], spatial navigation training [Lövdén et al., 2012], or intense studying of a foreign language [Mårtensson et al., 2012].

Automated morphometric methods, such as voxel-based morphometry [Ashburner and Friston, 2000], have been repeatedly criticized for inaccurate normalization procedures and for being susceptible to minor changes in the processing pipelines [Bookstein, 2001; Thomas et al., 2009; but see Ashburner and Friston, 2001, for a rebuttal]. Therefore, manual segmentation is still considered the gold standard. However, manual segmentation is time-consuming and labor-intensive and thus often not feasible for large data sets. It also requires at least two individual tracers to avoid biases. In many cases, it is desirable to segment the hippocampal formation quickly and efficiently, either to gather information on the health status of a specific person or to investigate large data sets resource-efficiently. Thus, it is important to further validate the usefulness of available automatic tools and to specify under which conditions they do and do not work well.

Manual and automatic segmentation methods have been compared previously [e.g., Cherbuin et al., 2009; Morey et al., 2009]. The findings mostly validate the use of automatic methods for segmentation of brain structures such as the Hc. Cherbuin et al. [2009] segmented 430 randomly selected individuals aged 44–48 years and showed that absolute hippocampal volumes were significantly larger with the automated measure compared with manual segmentation, supposedly due to relatively uniform over-inclusion of boundary voxels and surrounding cerebrospinal fluid, and the inclusion of the subiculum/entorhinal/parahippocampal regions. Still, correlations between the two methods were high and the relationship of hippocampal volume to selected sociodemographic and cognitive variables were unaffected by measurement method. Dewey et al. [2010] analyzed scans of 120 HIV-infected patients (86.7% male; with a mean age of 47.3 years) and compared two automated volumetric outputs, namely FreeSurfer and individual brain atlases using statistical parametric mapping (IBASPM), to auto-assisted manual tracings. They evaluated FreeSurfer to be effective for subcortical volumetry, but recommend visual inspection of segmentation output along with manual correction to ensure validity of the data. They suspected especially the border between amygdala and Hc to often be erroneous, as well as the border between tail of the Hc and lateral ventricle. Morey et al. [2009] compared manual tracings to automatically segmented hippocampal volumes with FreeSurfer and FSL/FIRST in 20 subjects. Validated in relation to hand tracing, hippocampal measurements with FreeSurfer were superior to FIRST on all aspects of the objective measures they used. They also reported a systematic inflation of Hc volume in Freesurfer estimates compared to hand tracing, which they attribute to shape differences in the anterior-medial surface and increased variance in this region of the Hc. Shen et al. [2010] compared manual and automated estimates of hippocampal volumes in patients with cognitive complaints (n = 39, mean age = 72.8), MCI (n = 37, mean age = 72.7), and early

AD (n = 11, mean age = 75.6) as well as in healthy controls (n = 38, mean age = 70.6). The two methods agreed strongly, confirming FreeSurfer's potential to determine hippocampal volumes in large-scale studies, even though there was a systematic volume difference between FreeSurfer and manual results. The authors argue that this difference stems from FreeSurfer's tendency to be more inclusive especially in the tail region and to have a few local excursions on the surfaces. Tae et al. [2008]) conducted a study on 21 female patients aged 18–60 years with major depressive disorder in which FreeSurfer showed good agreement with manual hippocampal volume and detected hippocampal atrophy in the patients compared with healthy controls. Similar to the studies mentioned earlier, the authors speculated that the volume overestimation was caused by an expansion of the Hc segmentation into adjacent white matter at the bottom of the Hc, as well as the inclusion of the entorhinal cortex and parts of the lateral ventricle, and an overall inaccurate differentiation of the Hc from the amygdala. Oscar-Berman and Song [2011] investigated the usefulness of FreeSurfer in 32 alcoholics and 37 controls with a mean age of about 53 years and showed that many brain structures as well as the Hc could be segmented reliably, but correlations between FreeSurfer and a semiautomated-supervised-system were lowest in subjects with the greatest abnormalities. Sánchez-Benavides et al. [2010] compared hippocampal automated measures with manual tracings in healthy (n = 41), mild cognitive impairment (MCI; n = 23), and Alzheimer's disease patients (n = 25; covering the age range of 60–80 years) and found adequate validity for FreeSurfer estimates with a general tendency for volume overestimation, even more pronounced in atrophic brains.

In this study, we compare results from a manual segmentation protocol that has extensively been described in Lövdén et al. [2012], mainly following the guidelines by Pruessner et al. [2000], to an automated segmentation of the hippocampus done with the FreeSurfer toolbox (www.surfer.nmr.mgh.harvard.edu). Here, we focus on two aspects that have not received sufficient attention so far. First, we investigate potential age group differences in reliability and validity of the automated volume estimation method of FreeSurfer. As an aged brain could also be seen as a brain with abnormalities that are more or less pronounced, depending on interindividual differences in biological aging, it is warranted to examine potential limitations of FreeSurfer in this population when comparing different age groups. Second, we investigate potential differences with respect to the segmentation of left and right Hc. Brain morphometric studies often incorporate hemispheric asymmetry analyses of segmented brain structures. Such asymmetry has been studied in many neurological conditions, for example post-traumatic stress disorder [Pavic et al., 2007], MCI and Alzheimer's disease [Chetelat and Baron, 2003; Geroldi et al., 2000; Shi et al., 2009], and depression [Kronmüller et al., 2009]. Since an analysis of hippocampal asymmetry should be important not only for diagnostic purposes, it is warranted to investigate the performance of automatic segmentation methods compared to manual tracing more closely also in normal populations.

METHODS

Participants and Study Design

To compare manual and automated segmentation of Hc, we used a previously acquired dataset including 44 younger participants (20–30 years; M = 26.0; SD = 2.8) and 47 older participants (60–70 years; M = 65.0; SD = 2.8). A detailed description of the effective sample and the design of the study have been reported elsewhere [see Lövdén et al., 2011, 2012; Wenger et al., 2012]. In short, the study consisted of a pretest measure (henceforth called time point A) with magnetic resonance imaging (MRI), a training phase of about 4 months, posttest 1 (henceforth called time point B) with MRI immediately after the training phase, a period of another 4 months without any training, followed by posttest 2 (henceforth called time point C) with a final MRI measure. Participants in the experimental group performed a navigation task in a virtual environment of a zoo while walking on a treadmill. Participants in the control group walked the exact same amount of time on a treadmill but without performing the navigation task. Here, we analyze participants from both groups together and the data from all three time points. Since the results for each time point are highly similar, we will focus on time point A. The replicated results from time point B and C can be found in the appendix. For cross-time point analyses, we use training group as a covariate to control for differential training effects across group and time, which have already been reported elsewhere [Lövdén et al., 2011, 2012; Wenger et al., 2012].

MRI Acquisition

Three high-resolution T_1-weighted images from each participant were acquired on the same 3 Tesla Magnetom Trio scanner (Siemens, Erlangen, Germany), with an 8-channel phased-array head coil. We used an MPRAGE sequence with the following parameters: TE = 5.12 ms, TR = 2600 ms, TI = 1100 ms, flip angle = 7°, bandwidth = 140 Hz/pixel, matrix = 320 × 320 × 240, isometric voxel size = 0.8 mm^3.

Manual Tracing

All manual tracing was performed based on the T_1-weighted MPRAGE images using a stylus on a Wacom DTU-710 pen tablet (Wacom Technology, Vancouver, WA) and the Analyze 8.1 software package (AnalyzeDirect, Overland Park, KS).

The two experienced raters were always blinded to group and measurement occasion before performing segmentation. Images were displayed in native space at a magnification factor of 3. We manually aligned all images and determined the anterior and posterior limits of the hippocampus to define the set of relevant slices. Both raters worked on a subset of randomly chosen slices of the same size, to minimize rater-related biases in single-data sets [Raz et al., 2004]. To enhance intraperson reliability across the 3 measurement occasions, we used a 2-step segmentation procedure. In the first step, one data set from each participant was randomly chosen and segmented according to the protocol described below. The resulting regions of interest (ROIs) were then coregistered to the remaining two data sets of that person, using automatic rigid-body coregistration routines provided in Analyze 8.1. All data sets were then anonymized with respect to the measurement point, such that the raters were unaware of the origin of the template ROI. The final segmentation was done on the basis of both the segmentation protocol and the template that overlaid the images from each time point. The entire Hc was segmented anterior-to-posterior using the coronal view as default. Generally, our approach followed the protocol provided by Pruessner et al. [2000] and was based on intrinsic anatomic properties of the hippocampus [Duvernoy, 2005] following a number of guiding rules. Full details of our protocol are reported as Supporting Information. To determine the reliability of the ROIs that entered our data analysis, we randomly chose a subsample of 20 data sets after the segmentation of the whole sample was completed. A parallel version of the ROIs that already existed for these data sets was added, so that each rater now processed those slices that the other rater had processed before. The interrater agreements for all slice-based and volume-based comparisons exceeded 0.93 [ICC2, Shrout and Fleiss, 1979].

FreeSurfer Segmentation

All cortical reconstruction and volumetric segmentation were performed with the FreeSurfer software package version 5.3 using the longitudinal processing scheme implemented to incorporate the subject-wise correlation of longitudinal data into the processing stream (http://surfer.nmr.mgh.harvard.edu/; e.g., Fischl et al., 2002) on a homogenous, GNU/Linux based computer cluster comprising of 608 CPU cores and a total memory size of 2.3 TB. Assessments of test-retest reliability of FreeSurfer have revealed high intraclass correlations of 0.994 for MPRAGE sequences [Wonderlick et al., 2009]. All reconstructed data were visually checked for segmentation accuracy at each time point. No manual interventions on the data were performed.

Statistical Analysis

All statistical analyses were conducted in SPSS 20, with an alpha level for all analyses of $p = 0.05$. We calculated stability coefficients over time for the two methods, with partial Pearson correlations using training group as a covariate. These test-retest correlations represent lower bound estimates of retest-reliability rather than test-retest reliability per se, given that they are calculated on a data set in which not only measurement error can introduce variance but also both normal aging and training effects can allow for true variance in change. All other analyses were run using the data from the first time point A. Replication of the reported pattern of results on the data of time point B and time point C can be found in the appendix.

We calculated bivariate Pearson correlations as well as intra-class correlation coefficients (ICC2) between the two estimation methods. Both measures can be used to compare two raters or methods. The Pearson correlation is a measure of interclass agreement that represents the overlap in relative ordering of individual data points in two data sets. In contrast, ICC are based on the assumption that two data sets pertain to one class with common variance and mean [Shrout and Fleiss, 1979; McGraw and Wong, 1996]. As a consequence, ICCs of the ratings of two methods do not only decrease in response to relative ordering of the single data points but also in response to differences in variance or differences in mean. In that sense, ICCs provide more stringent information about rater agreement [Müller and Büttner, 1994]. Conceptually, ICCs are more appropriate to quantify the agreement between manual and automated segmentation procedures. To provide better comparability to results of earlier studies [Cherbuin et al., 2009; Morrey et al., 2009; Sanchez-Benavides et al., 2010], however, we also report Pearson correlations.

As Shrout and Fleiss [1979] point out in their seminal article, there are many different types of ICC, and the version used must be selected carefully. Following suggestions by McGraw and Wong [1996], we used a two-way, mixed model, single measure intraclass correlation with agreement type, as is appropriate for the data at hand.

To investigate potential age and hemisphere differences, we conducted a 3-way repeated measures analysis of variance (ANOVA) with hemisphere (left, right) and method (manual, FreeSurfer) as within-subjects factors and age (young, old) as a between-subjects factor. To trace the source of significant interactions, we calculated 2-way repeated measures ANOVAs for each method, with hemisphere and age as factors. In addition, we conducted paired-samples t-tests to assess differences between estimation methods separately for the two hemispheres, and between hemispheres, for each estimation method. We conducted these follow-up analyses across as well as within age groups. To assess potential bias in the automatic volume estimation, we computed bivariate Pearson's correlations between manual estimates and the difference between FreeSurfer and manual estimates.

RESULTS

Stability Coefficients Over Time

Stability coefficients are summarized in Table I. Generally, partial correlations controlling for training group

TABLE I. Stability coefficients over time, separate for age groups, controlling for training group

	$r_{\text{pre, post1}}$		$r_{\text{post1, post2}}$		$r_{\text{pre, post2}}$	
	Left	Right	Left	Right	Left	Right
Young						
Manual	0.973	0.978	0.987	0.983	0.977	0.976
FreeSurfer	0.987	0.991	0.992	0.994	0.987	0.990
Old						
Manual	0.974	0.982	0.978	0.982	0.979	0.980
FreeSurfer	0.967	0.976	0.970	0.978	0.977	0.981

TABLE II. Bivariate Pearson correlations and intra-class correlations (agreement type) between manual segmentations and FreeSurfer estimates at time point A

	$r_{\text{manual, FreeSurfer}}$	Partial r controlling for age	r_{young}	r_{old}
Left	0.688	0.751	0.824	0.659
Right	0.785	0.836	0.890	0.784
	$\text{ICC}_{\text{overall}}$	$\text{ICC}_{\text{young}}$	ICC_{old}	
Left	0.321	0.275	0.361	
Right	0.466	0.381	0.577	

between the 3 measurement points were very high. For manual tracing, correlations were comparably high for younger and older adults alike, with no significant differences between correlations. For FreeSurfer segmentations, however, we observed significant age-group differences in the correlations. Correlations were lower for the older than for the younger age group between some time points: Between pretest and posttest 1, and posttest 1 and posttest 2, correlations were significantly lower in the older age group (computed with Fisher r-to-z transformation, $z \geq 2.27$, $p < 0.05$; see Table I, for exact correlation values).

Correlation Between FreeSurfer and Manual Volumes

Results for Pearson correlations between manual and FreeSurfer estimates as well as ICC values are listed in Table II. Pearson correlations between manual and FreeSurfer segmentations were .688 for the left Hc and .785 for the right Hc. Correlations between manual and FreeSurfer segmentations were numerically, though not significantly, higher in the young ($r_{\text{left}} = 0.824$, $r_{\text{right}} = 0.890$) than in the old ($r_{\text{left}} = 0.659$, $r_{\text{right}} = 0.784$). ICC estimates were considerably lower than Pearson correlations: 0.321 for the left Hc and 0.466 for the right Hc. In addition, age group comparisons indicated that the numerical pattern of agreement was reversed compared with Pearson correlations. ICCs were lower in the younger group ($\text{ICC}_{\text{left}} = 0.275$, $\text{ICC}_{\text{right}} = 0.381$) than in the older group ($\text{ICC}_{\text{left}} = 0.361$, $\text{ICC}_{\text{right}} = 0.577$), reflecting a more pronounced disagreement between estimates from manual and automated segmentation in the younger age group.

Mean Differences Between FreeSurfer and Manual Volumes as a Function of Age and Hemisphere

The 3-way ANOVA with Hemisphere, Method, and Age as factors revealed a significant main effect of Method, $F(1,89) = 565.39$, $p < 0.001$, $\eta_p^2 = 0.864$, indicating a consistently higher volume estimation in FreeSurfer, $t_{\text{left}}(90) = 17.12$, $p < 0.001$, $r^2 = 0.765$, $t_{\text{right}}(90) = 15.84$, $p < 0.001$, $r^2 = 0.736$ (see Table III for means and standard deviations). Further, we obtained a significant Method × Age Group interaction, $F(1,89) = 74.93$, $p < 0.001$, $\eta_p^2 = 0.457$, reflecting two facts: First, there was a larger mean volume difference in younger, $t_{\text{left}}(43) = 21.00$, $p < 0.001$, $r^2 = 0.911$, $t_{\text{right}}(43) = 21.31$, $p < 0.001$, $r^2 = 0.914$, than in older adults, $t_{\text{left}}(46) = 10.49$, $p < 0.001$, $r^2 = .705$, $t_{\text{right}}(46) = 8.80$, $p < .001$, $r^2 = .627$. Second, there was a significant age-related difference in hippocampal volume for FreeSurfer, but not for manual segmentation (Fig. 1). Hence, the main effect of Age in the 2-way ANOVA was reliable for FreeSurfer ($F(1,89) = 40.57$, $p < 0.001$, $\eta_p^2 = 0.313$), but not for manual tracing ($F(1,89) = 1.94$, $p = 0.168$, $\eta_p^2 = 0.021$).

The significant Hemisphere × Method interaction in the 3-way ANOVA, $F(1,89) = 25.47$, $p < 0.001$, $\eta_p^2 = 0.223$, indicates differences in hemispheric segmentation between the two methods (see Fig. 2). Specifically, with manual

TABLE III. Means and standard deviations of manual and FreeSurfer estimates, as well as absolute differences between the two methods at first measurement time point A

	Left			Right		
	Manual M ± SD	FreeSurfer M ± SD	Difference	Manual M ± SD	FreeSurfer M ± SD	Difference
All (n = 91)	3530.14 ± 379.31	4275.72 ± 492.54	745.58	3714.03 ± 413.50	4244.82 ± 515.14	530.579
Young (n = 44)	3582.56 ± 400.77	4464.69 ± 491.17	882.13	3778.16 ± 416.27	4520.15 ± 501.762	741.99
Old (n = 47)	3481.08 ± 355.35	3905.19 ± 307.38	424.11	3653.98 ± 406.12	3987.06 ± 378.93	333.08

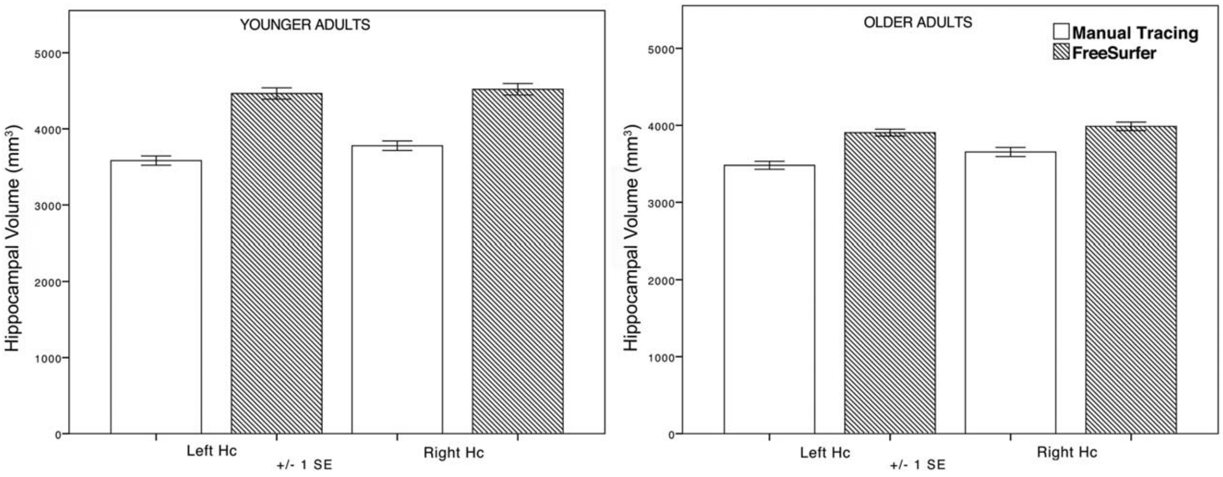

Figure 1.
Mean differences between the two estimation methods were more pronounced in the young age group.

segmentation, right Hc was estimated to be significantly larger than left Hc, $t(90) = 10.93$, $p < 0.001$, $r^2 = 0.570$. FreeSurfer also estimated the right Hc slightly larger than the left, however with a considerably smaller effect size, $t(90) = 2.86$, $p = 0.005$, $r^2 = 0.083$.

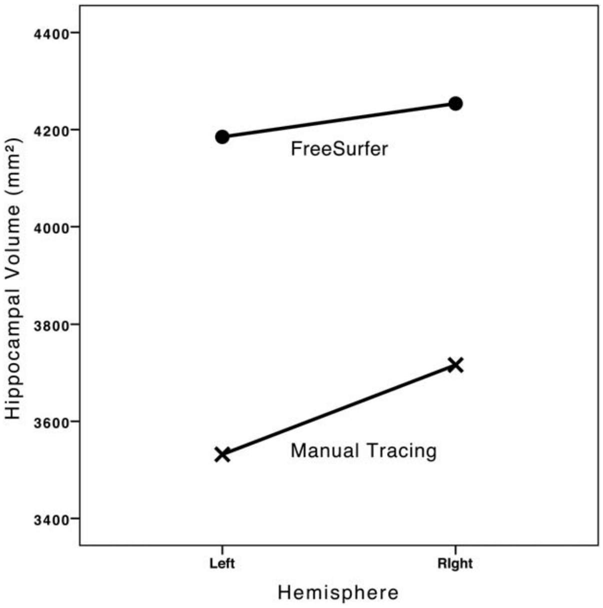

Figure 2.
Manual segmentation yielded a pronounced hemispheric difference in hippocampal volume. FreeSurfer also segmented the right Hc slightly larger than the left, however, with a considerably smaller effect size, reflected in a significant Hemisphere × Method interaction.

Correlation of Difference Between FreeSurfer and Manual Estimates and Manual Segmentation as a Proxy for True Hippocampal Volume

In the younger age group, there was no relation between manually assessed hippocampal volume and the difference between FreeSurfer and manual estimates, neither in left Hc, $r = 0.013$, $p = 0.931$, nor in right Hc, $r = 0.130$, $p = 0.401$. However, in the older age group there was a negative association, both in left Hc, $r = -0.551$, $p < 0.001$, and in right Hc, $r = -0.421$, $p = 0.003$. This suggests a bias in automatic volume estimation for the old age group. Specifically, in the younger age group, FreeSurfer overestimated hippocampal volume, and it did so independently of manually assessed hippocampal volume (Fig. 3; see also Fig. 4 for exemplary visualization of small and large young and old Hc segmentations). By contrast, in the older age group the difference between FreeSurfer and manual segmentation decreased as the manually segmented Hc size increased, indicating a differential bias toward relatively underestimating larger volumes in older adults.

The same analyses were run on the second and third measurement time point, where results display the exact same pattern as on the first time point (see appendix).

DISCUSSION

Our results reveal high stability coefficients over time for both manual and FreeSurfer segmentations. With FreeSurfer, correlations over time were significantly lower in the older than in the younger age group, which was not the case with manual segmentation. The Pearson correlations between the two assessment procedures were relatively high (0.683 for left Hc and 0.766 for right Hc).

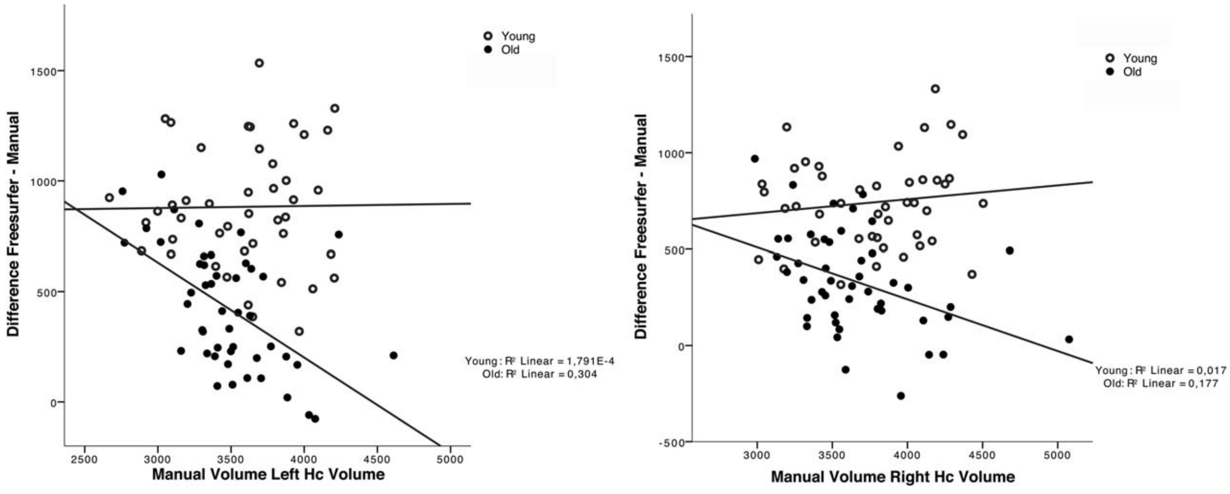

Figure 3.
Difference between FreeSurfer and manual estimates as a function of manually segmented "true" hippocampal volume decreases in the old age group but stays stable for the young age group.

Absolute agreements between the two measures, however, were considerably lower, as FreeSurfer estimated volumes to be higher. This volume difference was larger in the young than in the old. FreeSurfer detected a significant age difference in hippocampal volume, whereas manual tracing did not. However, manual tracing resulted in a significant difference between left and right Hc, whereas FreeSurfer segmented both sides in a more similar manner. FreeSurfer overestimated hippocampal size independently of manually assessed volume in the younger age group, but relatively underestimated larger volumes in older adults, thus introducing a bias in this age group.

To our knowledge, no other investigation up to date has compared manual segmentation with FreeSurfer segmentation in healthy younger adults and healthy older adults within the same study. Doing so has allowed us to discover an age-differential segmentation bias that would have gone unnoticed otherwise. For younger adults, individual differences in volume estimates derived from FreeSurfer and manual segmentation were highly correlated, as the mean volumes estimates with FreeSurfer exceeded manually obtained estimates by a constant amount. In contrast, for older adults, adding a constant value did not capture differences between the two methods, as the degree of disparity between the two methods was found to depend on the size of manually estimated volumes. Specifically, the difference between FreeSurfer and manual segmentation decreased as the manually segmented hippocampus volumes increased, pointing to a differential bias towards relatively underestimating larger volumes in older adults with FreeSurfer compared to manual segmentation. Thus, FreeSurfer may yield cross-sectional age-group differences in instances in which manual segmentation would not. This finding is novel and important for the field, as hippocampal segmentation based on FreeSurfer is increasingly being used in studies on normal cognitive aging, pathological cognitive aging, and in studies involving elderly populations suffering from some kind of disease.

These results thus portray a twofold picture of reliability and validity in manual and automatic Hc segmentations. In the younger age group, FreeSurfer can be regarded as a reliable and valid method for assessing differences in hippocampal volume, with at least as high reliability as manual tracing. However, in the older age group, the overestimation with FreeSurfer was dependent on "true" hippocampal volume, as it was smaller for larger hippocampi than for smaller hippocampi. Clearly, these novel results concerning an age-differential bias in older hippocampi are important for the growing field using automatic hippocampal segmentations in aging and diseased populations.

The absence of a significant main effect of age for our manual segmentation method may be regarded as surprising, given previous reports of hippocampal shrinkage in normal aging [e.g., Raz et al., 2005; Scahill et al., 2003]. However, there have also been studies finding no cross-sectional differences in hippocampal volume with age [Bigler et al., 1997; Du et al., 2006; Liu et al., 2003; Sullivan et al., 1995, 2005; Szentkuti et al., 2004]. It has also been shown previously that several individuals from an older age group in their seventies can indeed have a similar or even larger hippocampal volume than 20–30 year old adults, and that the dispersion around the mean of hippocampal volume does not necessarily have to differ for younger and older adults [Lupien et al., 2007], which is exactly what we see in our data from manual tracing. Furthermore, it is of course important to note the strict inclusion criteria that had to be met in order to participate in the study. Participants had to be able to walk on a treadmill without any gait problems, be free from neurological

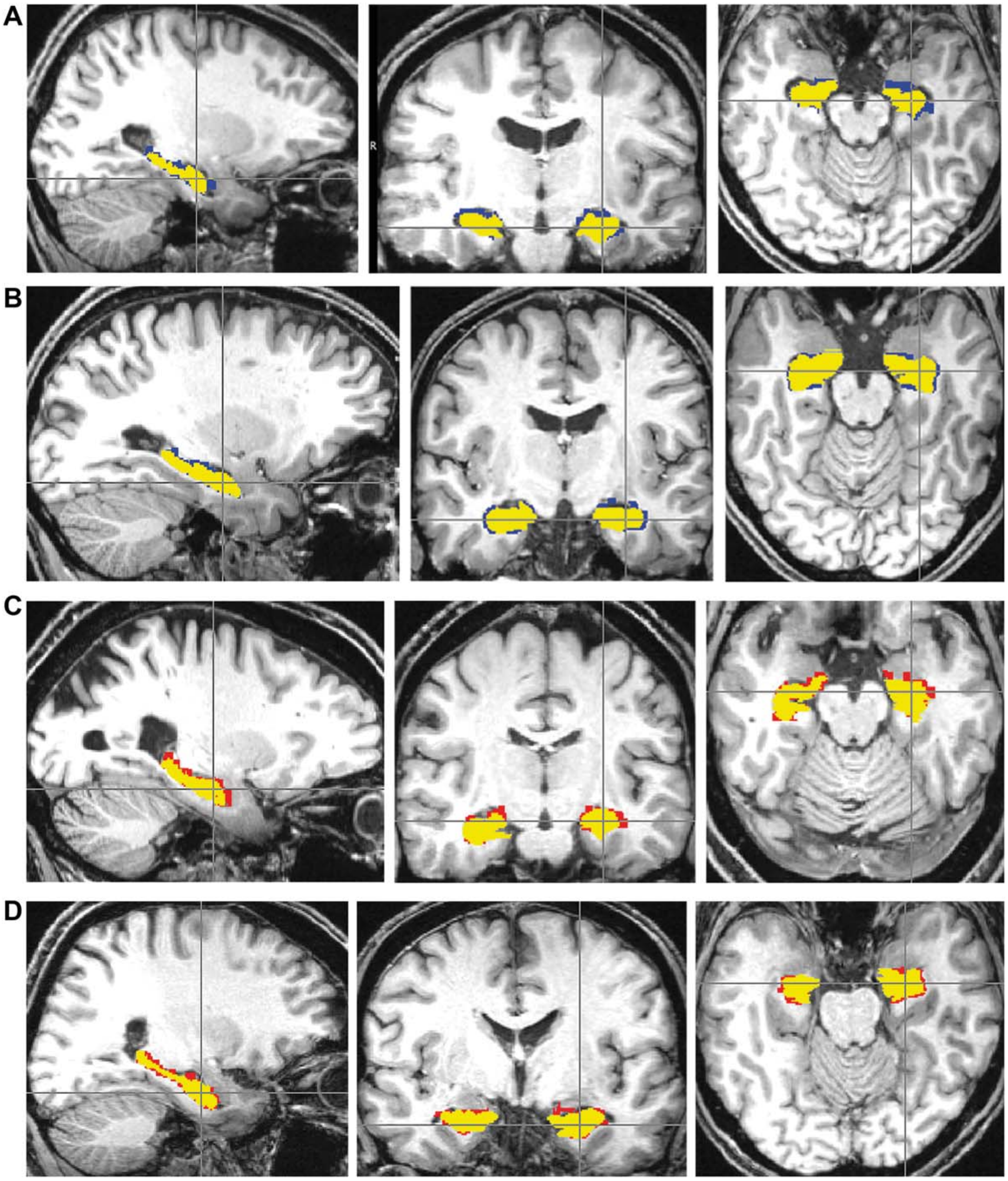

Figure 4.
A: Small Hc of a young participant. **B**: Large Hc of a young participant. **C**: Small Hc of an old participant. **D**: Large Hc of an old participant.

or other severe diseases, and had to be able to undergo MRI scanning, which removed participants with implants. Therefore, participants in our older age group may be classified as high-functioning elderly adults. As such, they are more likely to have a better-preserved brain structure than the average person from this age cohort.

The bias in volume estimation within FreeSurfer in the older age group is not only interfering with cross-sectional comparisons but also poses challenges on the programs' applicability in longitudinal studies. FreeSurfer would presumably be less sensitive to longitudinal decreases in hippocampal volume in samples of older adults, as the overestimation is higher for smaller volumes in this age group, and would also be less likely to find longitudinal increases, as larger hippocampal volumes are not being overestimated to the same extent as smaller ones.

Another intriguing difference between the two segmentation methods concerns the left–right Hc asymmetry. As expected from the literature [e.g., Barnes et al., 2008; Basso et al., 2006; Honeycutt and Smith, 1995; Jack et al., 2003; Krasuski et al., 1998; Pruessner et al., 2000], manual segmentation yielded a bigger right than left Hc. FreeSurfer also estimates the right Hc as slightly bigger, but with a considerably smaller effect size. When we used earlier processing versions of FreeSurfer, we even saw an absence of significant differences between the hippocampi. The data pattern observed in our study is in good agreement with the left-right asymmetry reported in Sánchez-Benavides et al. [2010]. Sánchez-Benavides et al. [2010] found that the left-right discrepancy was statistically reliable with manual segmentation, but not with FreeSurfer segmentation. Although the left-right asymmetric segmentations result is repeatedly reported in several studies, there are also others that do not find a significant difference between left and right hippocampus in healthy individuals, both with segmentation on MR images [e.g., Bigler et al., 1997; Raz et al., 2004] and in autopsy data of post-mortem brains [Bogerts et al., 1990]. This ambiguity might stem from a potential bias in manual segmentation, based on the laterality of visual perception [Maltbie et al., 2012]. It is possible that human tracers do prefer one side of a displayed brain during tracing and are thus more accurate on one side of the brain, therefore producing one bigger and one smaller Hc. A possible solution could be to display both hippocampi on the same side of the monitor by flipping the MR image during segmentation. As it cannot be determined from manual segmentation results alone how big the true anatomical left-right asymmetry effect is, it cannot be concluded whether FreeSurfer is performing erroneously or correctly in producing a considerably smaller effect.

We have discussed that FreeSurfer's reliability and validity is different for younger and older hippocampi. It is quite likely that the values for tissue priors and coordinates of the default segmentation atlas used in FreeSurfer are further away from optimal values for older adults than for younger adults. The default segmentation atlas in FreeSurfer is created on the basis of 39 middle-aged brains (mean age around 38 years ± 10), primarily drawn from a sample described by Goldstein et al. [1999]. The hippocampi of this sample were manually segmented according to the conventions of the Center for Morphometric Analysis [Fischl et al., 2002, 2004]. Probably, the brain structure of these middle-aged brains was more similar to the brain structure of the younger adults of our study, who were aged 20–30 years, than to the brain structure of the older adults, who were aged 60–70 years, resulting in more accurate segmentation for younger than for older brains. Older brain structures, as displayed on T_1-weighted MR images, typically show a less distinct gray-white matter contrast [Westlye et al., 2009], and are more "blurry," which makes it harder to define the true border around the Hc and its surrounding structures or cerebrospinal fluid. However, it remains unknown why the automatic segmentation algorithm did not overestimate the volume for bigger older Hc in the same way as for smaller older Hc, or younger Hc in general. Considering the blurriness of old brain structures on MR images, it would have been maybe even more plausible if bigger older Hc had been more overestimated. However, the present data yielded the opposite pattern. Future studies should follow up on this finding of overestimation in younger and smaller older Hc, coupled with a relative underestimation of larger older Hc and try to determine sources of this estimation bias. A useful step would be to build age group-specific default segmentation atlases [see also Avants et al., 2010]. Given that normal aging is associated with gradual anatomical changes [e.g., Raz et al., 2005] and is often modulated by age-correlated conditions that affect the brain, such as cardiovascular and metabolic syndromes, the importance of arriving at an age-adjusted template is obvious. Rerunning the FreeSurfer pipeline with customized masks in the background and examining how such default-segmentation atlases may change subcortical volume estimates may help to circumvent the current ambiguities in the data. As the generation of a whole-brain segmentation atlas is very time-consuming, cooperation between different laboratories performing manual whole brain segmentation is desirable, with the common goal of creating an age-graded segmentation atlas.

In general, there are marked differences in the labeling protocol between FreeSurfer and our manual segmentations, which followed the rules provided by Pruessner et al. [2000]. Visual inspection of the graphic overlay of both segmentations gives hints to differences in the labeling protocol especially in the subiculum/entorhinal/parahippocampal regions, the border between amygdala and the hippocampus, and the border between the tail of the hippocampus and the lateral ventricle, as has also been reported by others [e.g., Cherbuin et al., 2009; Dewey et al., 2010; Morey et al., 2009]. In all these areas, FreeSurfer seems to be more inclusive, whereas the manual tracing protocol provided clear rules of where to draw the anatomical boundary between the hippocampus and its surroundings, which is in general difficult to define [Duvernoy, 2005; also see Supporting Information for further information]. These differences in segmentation protocols might prevent the two methods from being completely interchangeable, and might pose challenges on studies in which hippocampal shape in these regions is of main interest. The tendency of FreeSurfer towards overinclusion of boundary voxels and a more liberal definition of boundaries to the amygdala, entorhinal cortex, and lateral ventricle would not constitute a major drawback for many research questions if it were consistent across individuals and volumes. However, the results of this study show that this tendency varies by volume and age group, resulting in biased comparisons between younger and older adults, and biased individual estimates within groups of older adults. We note a need for further improvements of the FreeSurfer segmentation pipeline to identify and overcome these sources of bias.

To summarize, there were differences in segmentation outcomes between manual and FreeSurfer estimates with regard to age effects and left-right asymmetry, with further investigation needed to establish which segmentation method introduces a bias in volume estimation and left-right asymmetry. We conclude that FreeSurfer is a credible and valid method to assess differences in hippocampal volume in younger age groups and can therefore be a feasible method to analyze large datasets of young adults. However, no definitive statements about the size of left and right Hc should be made at this point, and research where such asymmetries are of importance should take heed. Importantly, FreeSurfer estimates in older age groups should be interpreted with care until the automatic segmentation pipeline has been further optimized to increase validity and reliability in this age group.

ACKNOWLEDGMENTS

The authors thank Colin Bauer, Daniel Bittner, Marcel Derks, Isabel Dziobek, Gabriele Faust, Martin Kanowski, Jörn Kaufmann, Kristen Kennedy, Shireen Kwiatkowska-Naqvi, Naftali Raz, Karen Rodrigue, Michael Schellenbach, Claus Tempelmann, and all student assistants.

APPENDIX

Similar to the data from the first measurement time point, correlations between manual and FreeSurfer estimates were numerically higher in the younger than in the older age group at both later time points B and C (see Table IV). FreeSurfer again yielded consistently higher hippocampal volume than manual segmentation, which was again more pronounced in younger adults. Manual segmentation resulted in a hemispheric difference between left and right hippocampus, whereas FreeSurfer estimated them more similarly (see Table V). Replicating results from the first measurement time point, we found no relation between manually assessed hippocampal volume and the difference between manual and FreeSurfer estimates in younger adults (time point B: $r_{\text{left Hc}} = 0.073$, $p = 0.637$, $r_{\text{right Hc}} = 0.123$, $p = 0.428$; time point C: $r_{\text{left Hc}} = 0.032$, $p = 0.836$, $r_{\text{right Hc}} = 0.178$, $p = 0.247$), but a negative association in the older age group (time point B: $r_{\text{left Hc}} = -0.550$, $p < 0.001$, $r_{\text{right Hc}} = -0.465$, $p = 0.001$; time point C: $r_{\text{left Hc}} = -0.538$, $p < 0.001$, $r_{\text{right Hc}} = -0.447$, $p = 0.002$). Again, FreeSurfer overestimated hippocampal volume consistently in the young age group, but, in the older age group, FreeSurfer overestimated large hippocampi to a less extent than small hippocampi.

TABLE IV. Bivariate Pearson correlations and intraclass correlations between manual segmentations and FreeSurfer estimates at time point B and C

	$r_{\text{manual, FreeSurfer}}$	Partial r controlling for age	r_{young}	r_{old}
Time point B				
Left	0.695	0.746	0.829	0.641
Right	0.778	0.827	0.869	0.798
	$\text{ICC}_{\text{overall}}$	$\text{ICC}_{\text{young}}$	ICC_{old}	
Left	0.325	0.276	0.351	
Right	0.456	0.358	0.579	
Time point C				
Left	0.687	0.764	0.834	0.683
Right	0.765	0.814	0.873	0.768
	$\text{ICC}_{\text{overall}}$	$\text{ICC}_{\text{young}}$	ICC_{old}	
Left	0.323	0.266	0.397	
Right	0.452	0.356	0.574	

TABLE V. Means and standard deviations of manual and FreeSurfer estimates for timepoint B and C, as well as absolute differences between the two methods

	Left			Right		
Time point B	Manual M ± SD	FreeSurfer M ± SD	Difference	Manual M ± SD	FreeSurfer M ± SD	Difference
All (n = 91)	3519.88 ± 380.37	4171.21 ± 504.28	651.33	3698.67 ± 407.58	4233.30 ± 513.78	534.63
Young (n = 44)	3582.92 ± 395.41	4467.50 ± 501.65	884.58	3771.19 ± 401.50	4520.83 ± 497.03	749.64
Old (n = 47)	3460.86 ± 359.93	3892.89 ± 314.03	432.03	3630.77 ± 406.12	3964.12 ± 363.90	333.35
Time point C						
All (n = 91)	3527.08 ± 379.73	4171.00 ± 499.18	643.92	3695.11 ± 411.43	4232.42 ± 528.22	537.31
Young (n= 44)	3576.84 ± 393.43	4473.74 ± 482.06	896.90	3764.50 ± 398.81	4524.55 ± 508.03	760.05
Old (n = 47)	3480.50 ± 364.48	3887.58 ± 317.13	407.08	3630.15 ± 416.65	3958.93 ± 382.85	328.78

REFERENCES

Ashburner J, Friston KJ (2000): Voxel-based morphometry - The methods. Neuroimage 11:805-821. doi: http://dx.doi.org/10.1006/nimg.2000.0582.

Ashburner, J, Friston KJ (2001): Why voxel-based morphometry should be used. Neuroimage 14:1238-1243. doi: http://dx.doi.org/10.1006/nimg.2001.0961.

Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee JC (2010): The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49:2457-2466. doi: http://dx.doi.org/10.1016/j.neuroimage.2009.09.062.

Barnes J, Foster J, Boyes RG, Pepple T, Moore EK, Schott JM, Fox NC (2008): A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. Neuroimage 40:1655-1671. doi: http://dx.doi.org/10.1016/j.neuro image.2008.01.012.

Basso M, Yang J, Warren L, MacAvoy MG, Varma P, Bronen Ra, Van Dyck CH (2006): Volumetry of amygdala and hippocampus and memory performance in Alzheimer's disease. Psychiatry Res 146:251–261. doi:10.1016/j.pscychresns.2006.01.007.

Bhatia S, Bookheimer SY, Gaillard WD, Theodore WH (1993): Measurement of whole temporal lobe and hippocampus for MR volumetry: Normative data. Neurology 43:2006-2010. doi: http://dx.doi.org/10.1212/WNL.43.10.2006.

Bigler ED, Blatter DD, Anderson CV, Johnson SC, Gale SD, Hopkins RO, Burnett B (1997): Hippocampal volume in normal aging and traumatic brain injury. Am J Neuroradiol 18:11-23.

Bogerts B, Falkai P, Haupts M, Greve B, Ernst S, Tapernon-Franz U, Heinzmann U (1990): Post-mortem volume measurements of limbic system and basal ganglia structures in chronic schizophrenics: Initial results from a new brain collection. Schizophr Res 3:295-301. doi: http://dx.doi.org/10.1016/0920-9964(90)90013-W.

Bookstein FL (2001): "Voxel-based morphometry" should not be used with imperfectly registered images. Neuroimage 14:1454-1462. doi: http://dx.doi.org/10.1006/nimg.2001.0770.

Boyke J, Driemeyer J, Gaser C, Büchel C, May A (2008): Training-induced brain structure changes in the elderly. J Neurosci 28: 7031-7035. doi: http://dx.doi.org/10.1523/JNEUROSCI.0742-08.2008.

Bremner JD, Randall P, Scott TM, Bronen RA, Seibyl JP, Southwick SM, Innis RB (1995): MRI-based measurement of hippocampal volume in patients with combat-related postraumatic stress disorder. Am J Psychiatry 152:973-981.

Cherbuin N, Anstey KJ, Réglade-Meslin C, Sachdev PS (2009): In vivo hippocampal measurement and memory: A comparison of manual tracing and automated segmentation in a large community-based sample. PLoS One 4:e5265. doi: http://dx.doi.org/10.1371/journal.pone.0005265

Chetelat G, Baron JC (2003): Early diagnosis of Alzheimer's disease: Contribution of structural neuroimaging. NeuroImage 18: 525-541. doi: http://dx.doi.org/10.1016/S1053-8119(02)00026-5.

Christie BR, Cameron HA (2006): Neurogenesis in the adult hippocampus. Hippocampus 16:199-207. doi: 10.1002/hipo.20151.

Churchhill JD, Galvez R, Colcombe S, Swain RA, Kramer AF, Greenough WT (2002): Exercise, experience and the aging brain. Neurobiol Aging 23:941-955. doi: http://dx.doi.org/10.1016/S0197-4580(02)00028-3.

Dewey J, Hana G, Russell T, Price J, McCaffrey D, Harezlak J, Consortium HN (2010): Reliability and validity of MRI-based automated volumetry software realtive to auto-assisted manual measurement of subcortical structures in HIV-infected patients from a multisite study. Neuroimage 51:1334-1344. doi: http://dx.doi.org/10.1016/j.neuroimage.2010.03.033.

Draganski B, Gaser C, Kempermann G, Kuhn HG, Winkler J, Büchel C, May A (2006): Temporal and spatial dynamics of brain structure changes during extensive learning. J Neurosci 26:6314-6317. doi: http://dx.doi.org/10.1523/JNEUROSCI.46 28-05.2006.

Du AT, Schuff N, Chao LL, Kornak J, Jagust WJ, Kramer JH, Weiner MW (2006): Age effects on atrophy rates of entorhinal cortex and hippocampus. Neurobiol Aging 27:733-740. doi: http://dx.doi.org/10.1016/j.neurobiolaging.2005.03.021.

Duvernoy HM (2005): The Human Hippocampus: Functional Anatomy, Vascularization And Serial Sections with MRI, 3rd ed. Berlin: Springer-Verlag.

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, Dale AM (2002): Whole brain segmentation: Automated labeling of neuroanatomic structures in the human brain. Neuron 33:341-355. doi: http://dx.doi.org/10.1016/S0896-6273(02)0 0569-X.

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Dale AM (2004): Automatically Parcellating the human cerebral cortex. Cereb Cortex 14:11-22. doi: http://dx.doi.org/10.1093/cercor/bhg087.

Geroldi C, Laasko MP, DeCarli C, Beltamello A, Bianchetti A, Soininen H, Frisoni GB (2000): Apolipoprotein E genotype and hippocampal asymmetry in Alzheimer's disease: A volumetric MRI study. J Neurol Neurosurg Psychiatry 68:93-96. doi: http://dx.doi.org/10.1136/jnnp.68.1.93.

Goldstein JM, Goodman JM, Seidman LJ, Kennedy DN, Makris N, Lee H, Tsuang MT (1999): Cortical abnormalities in schizophrenia identified by structural magnetic resonance imaging. Arch General Psychiatry 56:537-547. doi: http://dx.doi.org/10.1001/archpsyc.56.6.537.

Heckers S (2001): Neuroimaging studies of the hippocamps in schozophrenia. Hippocampus 11:520-528. doi: http://dx.doi.org/10.1002/hipo.1068.

Honeycutt NA, Smith CD (1995): Hippocampal volume measurements using magnetic resonance imaging in normal young adults. J Neuroimaging 5:95-100.

Jack Jr CR, Peterson RC, Xu YC, O'Brian eC, Smith GE, Ivnik RJ, Kokmen E (1999): Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 52:1397-1403. doi: http://dx.doi.org/10.1212/WNL.52.7.1397.

Jack Jr CR, Slomkowski M, Gracon S, Hoover TM, Felmlee JP, Stewat K, Xu Y, et al (2003): MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. Neurology 60:253–260. doi: http://dx.doi.org/10.1212/01.WNL.0000042480.86872.03.

Jessberger S, Gage FH (2008): Stem cell-associated structural and functional plasticity in the aging hippocampus. Psychol Aging 23:684-691. doi: http://dx.doi.org/10.1037/a0014188.

Kempermann G, Gast D, Gage FH (2002): Neuroplasticity in old age: Sustained fivefold induction of hippocampal neurogenesis by long-term environmental enrichment. Ann Neurol 52:135-143. doi: http://dx.doi.org/10.1002/ana.10262.

Krasuski JS, Alexander GE, Horwitz B, Daly EM, Murphy DGM, Rapoport SI, Schapiro MB (1998): Volumes of medial temporal lobe structures in patients with Alzheimer's disease and mild cognitive impairment (and in Healthy Controls). Biol Psychiatry 43:60-68. doi: http://dx.doi.org/10.1016/S0006-3223(97)0 0013-9.

Kronenberg G, Bick-Sander A, Bunk E, Wolf C, Ehninger D, Kempermann G (2006): Physical exercise prevents age-related decline in precursor cell activity in the mouse dntate gyrus. Neurobiol Aging 27:1505-1513. doi: http://dx.doi.org/10.1016/j.neurobiolaging.2005.09.016.

Kronmüller KT, Schröder J, Köhler S, Götz B, Victor D, Unger J, Essig M, Pantel J (2009): Hippocampal volume in first episode and recurrent depression. Psychiatry Res Neuroimaging 174: 62-66. doi: http://dx.doi.org/10.1016/j.pscychresns.2008.08.001.

Liu RS, Lemieux L, Bell GS, Sisodiya SM, Shorvon SD, Sander JW, Duncan JS (2003): A longitudinal study of brain morphometrics using quantitative magnetic resonance imaging and difference image analysis. NeuroImage 20:22-33. doi: http://dx.doi.org/10.1016/S1053-8119(03)00219-2.

Lövdén M, Schaefer S, Noack H, Kanowski M, Kaufmann J, Tempelmann C, Bäckman L (2011): Performance-related increases in hippocampal N-acetylaspartate (NAA) induced by spatial navigation training are restricted to BDNF Val homozygotes. Cerebral Cortex 21:1435-1442.

Lövdén M, Schaefer S, Noack H, Bodammer NC, Kühn S, Heinze H-J, Lindenberger U (2012): Spatial navigation training protects the hippocampus against age-related changes during early and late adulthood. Neurobiol Aging 33:620.e629-620.e622. doi: http://dx.doi.org/10.1016/j.neurobiolaging.2011.02.013.

Lupien SJ, Evans A, Lord C, Miles J, Pruessner M, Pike B, Pruessner JC (2007): Hippocampal volume is as variable in young as in older adults: Implications for the notion of hippocamapl atrophy in humans. NeuroImage 34:479-485. doi: http://dx.doi.org/10.1016/j.neuroimage.2006.09.041.

Maguire EA, Gadian DG, Johnsrude IS, Good CD, Ashburner J, Frackowiak RS, Frith CD (2000): Navigation-related structural change in the hippocampus of taxi drivers. Proc Natl Acad Sci USA 97:4398-4403. doi: http://dx.doi.org/10.1073/pnas.070039597.

Maguire EA, Woollett K, Spiers HJ (2006): London taxi and bus drivers: A structural MRI and neuropsychological analysis. Hippocampus 16:1091-1101. doi: http://dx.doi.org/10.1002/hipo.20233.

Maltbie E, Bhatt K, Paniagua B, Smith RG, Graves MM, Mosconi MW, Styner MA (2012): Asymmetric bias in user guided segmentations of brain structures. Neuroimage 59:1315-1323. doi: http://dx.doi.org/10.1016/j.neuroimage.2011.08.025.

Mårtensson J, Eriksson J, Bodammer NC, Lindgren M, Johansson M, Lövdén M (2012): Growth of language-related brain areas after foreign language learning. NeuroImage 63:240-244. doi: dx.doi.org/10.1016/j.neuroimage.2012.06.043.

McGraw KO, Wong SP (1996): Forming inferences about some intraclass correlation coefficients. Psychol Methods 1:30–46. doi: http://dx.doi.org/10.1037/1082-989X.1.1.30.

Morey RA, Petty CM, Xu Y, Hayes JP, Wagner HR II, Lewis DV, McCarthy G (2009): A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45:855-866. doi: http://dx.doi.org/10.1016/j.neuroimage.2008.12.033.

Müller R, Büttner P (1994): A critical discussion of intraclass correlation coefficients. Stat Med 13:2465–2476. doi: http://dx.doi.org/10.1002/sim.4780132310.

Müller SG, Schuff N, Yaffe K, Madison C, Miller B, Weiner MW (2010): Hippocampal atrophy patterns in mild cognitive impairment and alzheimer's disease. Hum Brain Mapp 31: 1339-1347. doi: http://dx.doi.org/10.1002/hbm.20934.

Oscar-Berman M, Song J (2011): Brain volumetric measures in alcoholics: A comparison of two segmnetaiton methods. Neuropsychiatr Dis Treat 7:65-75. doi: http://dx.doi.org/10.2147/NDT.S13405.

Pavic L, Rudolf G, Rados M, Brklijacic B, Brajkovic L, Simentin-Pavic L, Kalousek V (2007): Smaler right hippocampus in war veterans with posttraumatic stress disorder. Psychiatry Res: Neuroimaging 154:191-198.

Pruessner JC, Li LM, Series W, Pruessner M, Collins DL, Kabani N, Evans AC (2000): Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: Minimizing the discrepancies between laboratories. Cerebral Cortex 10:433-442. doi: http://dx.doi.org/10.1093/cercor/10.4.433.

Raz N, Gunning-Dixon F, Head D, Rodrigue KM, Williamson A, Acker JD (2004): Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: Replicability of regional differences in volume. Neurobiol Aging 25:377-396. doi: http://dx.doi.org/10.1016/S0197-4580(03)00118-0.

Raz N, Lindenberger U, Rodrigue KM, Kennedy KM, Head D, Williamson A, Acker JD (2005): Regional brain changes in aging healthy adults: General trend, individual differences and modifiers. Cerebral Cortex 15:1676-1689.

Rosenzweig MR, Bennett EL (1996): Psychobiology of plasticity: Effects of training and experience on brain and behavior. Behav Brain Res 78:57-65. doi: http://dx.doi.org/10.1016/0166-4328(95)00216-2.

Sánchez-Benavides G, Gómez-Ansón B, Sainz A, Vives Y, Delfino M, Peña-Casanova J (2010): Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. Psychiatry Res 181:219–25. doi:10.1016/j.pscychresns.2009.10.011.

Scahill RI, Frost C, Jenkins R, Whitwell JL, Rossor MN, Fox NC (2003): A longitudinal study of brain volume changes in normal aging using serial magnetic resonance imaging. Arch Neurol 60:989-994. doi: http://dx.doi.org/10.1001/archneur.60.7.989.

Sheline Y, Wang pW, Gado MH, Csernansky JG, Vannier MW (1996): Hippocampal atrophy in recurrent major depression. Proc Natl Acad Sci USA 93:3908-3913. doi: http://dx.doi.org/10.1073/pnas.93.9.3908.

Shen L, Sayhin AJ, Kim S, Firpi iA, West JD, Risacher SL, Flashman LA (2010): Comparison of manual and automated determination of hippocampal volumes in MCI and early AD. Brain Imaging Behavior 4:86-95. doi: http://dx.doi.org/10.1007/s11682-010-9088-x.

Shi F, Liu B, Zhou Y, Yu C, Jiang T (2009): Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. Hippocampus 19:1055-1064. doi: http://dx.doi.org/10.1002/hipo.20573.

Shrout PE, Fleiss JL (1979): Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420-428. doi: http://dx.doi.org/10.1037//0033-2909.86.2.420.

Squire LR, Stark CEL, Clark RE (2004): The medial temporal lobe. Annu Rev Neurosci 27:279-306. doi: http://dx.doi.org/10.1146/annurev.neuro.27.070203.144130.

Sullivan EV, Marsh L, Mathalon DH, Lim KO, Pfefferbaum A (1995): Age-related decline in MRI volumes of temporal lobe gray matter but not hippocampus. Neurobiol Aging 16:591-606. doi: http://dx.doi.org/10.1016/0197-4580(95)00074-O.

Sullivan EV, Marsh L, Pfefferbaum A (2005): Preservation of hippocampal volume throughout adulthood in healthy men and women. Neurobiol Aging 26:1093-1098. doi: http://dx.doi.org/10.1016/j.neurobiolaging.2004.09.015.

Szentkuti A, Guderian S, Schiltz K, Kaufmann J, Münte TF, Heinze H-J, Düzel E (2004): Quantitative MR analyses of the hippocampus: Unspecific metabolic changes in aging. J Neurol 251:1345-1353. doi: 10.1007/s00415-004-0540-y.

Tae WS, Kim SS, Lee KU, Nam E-C, Kim KW (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. doi: http://dx.doi.org/10.1007/s00234-008-0383-9.

Thomas AG, Marrett S, Saad ZS, Ruff DA, Martin A, Bandettini PA (2009): Functional but not structural changes associated with learning: An exploration of longitudinal Voxel-Based Morphometry (VBM). Neuroimage 48:117-125. doi: http://dx.doi.org/10.1016/j.neuroimage.2009.05.097.

van Praag H, Kempermann G, Gage FH (2000): Neural consequences of environmental enrichment. Nat Rev Neurosci 1:191-198. doi: http://dx.doi.org/10.1038/35044558.

Wenger E, Schaefer S, Noack H, Kühn S, Martensson J, Heinze H-J, Lövdén M (2012): Cortical thickness changes following spatial navigation training in adulthood and aging. Neuroimage 59:3389-3397. doi: http://dx.doi.org/.

Westlye LT, Walhovd KB, Dale AM, Espeseth T, Reinvang I, Raz N, Fjell AM (2009): Increased sensitivity to effects of normal aging and Alzheimer's disease on cortical thickness by adjustment for local variability in gray/white contrast: A multisample MRI study. Neuroimage 47:1545-1557. doi: http://dx.doi.org/10.1016/j.neuroimage.2009.05.084.

Wonderlick JS, Ziegler DA, Hosseini-Varnamkhasti P, Locascio JJ, Bakkour A, van der Kouwe A, Dickerson BC (2009): Reliability of MRI-derived cortical and subcortical morphometric measures: Effects of pulse sequence, voxel geometry, and parallel imaging. Neuroimage 44:1324-1333. doi: http://dx.doi.org/10.1016/j.neuroimage.2008.10.037.

Woollett K, Maguire EA (2011): Acquiring "the knowledge" of London's layout drives structural brain changes. Curr Biol 21:2109-2114. doi: http://dx.doi.org/10.1016/j.cub.2011.11.018.