Received: 8 September 2017 | Revised: 11 February 2018 | Accepted: 13 February 2018

DOI: 10.1002/hbm.24017

RESEARCH ARTICLE

WILEY

A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan

Mike F Schmidt[1,2] | Judd M Storrs[3] | Kevin B Freeman[4] | Clifford R Jack Jr[5] | Stephen T Turner[6] | Michael E Griswold[7] | Thomas H Mosley Jr[2]

[1]Program in Neuroscience, University of Mississippi Medical Center, Jackson, Mississippi

[2]Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi

[3]Department of Radiology, University of Mississippi Medical Center, Jackson, Mississippi

[4]Department of Psychiatry and Human Behavior, University of Mississippi Medical Center, Jackson, Mississippi

[5]Department of Radiology, Mayo Clinic, Rochester, Minnesota

[6]Division of Nephrology and Hypertension, Mayo Clinic, Rochester, Minnesota

[7]Department of Data Science, University of Mississippi Medical Center, Jackson, Mississippi

Correspondence
Mike F Schmidt, 260 Park Ave, North Caldwell, NJ 07006.
Email: mikeschmidt@schmidtgracen.com

Funding information
National Heart, Lung, and Blood Institute, Grant/Award Number: U01HL054463; National Institute of Neurological Disorders and Stroke, Grant/Award Number: R01NS041558

Abstract

MRI has become an indispensable tool for brain volumetric studies, with the hippocampus an important region of interest. Automation of the MRI segmentation process has helped advance the field by facilitating the volumetric analysis of larger cohorts and more studies. FreeSurfer has emerged as the de facto standard tool for these analyses, but studies validating its output are all based on older versions. To characterize FreeSurfer's validity, we compare several versions of FreeSurfer software with traditional hand-tracing. Using MRI images of 262 males and 402 females aged 38 to 84, we directly compare estimates of hippocampal volume from multiple versions of FreeSurfer, its hippocampal subfield routines, and our manual tracing protocol. We then use those estimates to assess asymmetry and atrophy, comparing performance of different estimators with each other and with brain atrophy measures. FreeSurfer consistently reports larger volumes than manual tracing. This difference is smaller in larger hippocampi or older people, with these biases weaker in version 6.0.0 than prior versions. All methods tested agree qualitatively on rightward asymmetry and increasing atrophy in older people. FreeSurfer saves time and money, and approximates the same atrophy measures as manual tracing, but it introduces biases that could require statistical adjustments in some studies.

KEYWORDS

aging, asymmetry, FreeSurfer, hippocampus, human, magnetic resonance imaging, segmentation

1 | INTRODUCTION

The hippocampus is a part of the medial temporal lobe, playing an important role in spatial and episodic memory (Squire, Stark, & Clark, 2004). The ability to measure hippocampal volume in humans via magnetic resonance imaging (MRI) has identified reduced hippocampal volume as a biomarker for numerous conditions (Geuze, Vermetten, & Bremner, 2005), including major depressive disorder (MDD; Campbell, Marriott, Nahmias, & MacQueen, 2004; Videbech & Ravnkilde, 2004), epilepsy (Cook, Fish, Shorvon, Straughan, & Stevens, 1992; Geuze et al., 2005; Jack et al., 1990; Jack, 1994), post-traumatic stress disorder (Logue et al., 2017), and Alzheimer disease (AD; Gosche, Mortimer, Smith, Markesbery, & Snowdon, 2002; Jack et al., 1998; Kesslak, Nalcioglu, & Cotman, 1991), as well as normal aging (Allen, Bruss, Brown, & Damasio, 2005; Du et al., 2006; Raz, Rodrigue, Head, Kennedy, & Acker, 2004b; Raz & Rodrigue, 2006; Walhovd et al., 2005; Walhovd et al., 2011). Increasing MRI availability and concurrent reductions in costs have led to a proliferation of MRI data in ever larger studies and data sets.

Determining hippocampal volume from MRI has been accomplished through varied manual (Boccardi et al., 2011; Jack, 1994; Konrad et al., 2009) and automated (Alemán-Gómez, Melie-Garcia, & Valdes-Hernandez, 2006; Fischl et al., 2002; Moghaddam & Soltanian-Zadeh, 2009; Platero & Tobar 2016; Smith et al., 2004) methods. Manual tracing has long been considered the gold standard, allowing for flexible quantification without many of the assumptions built into algorithms. Experienced human tracers can correctly mark ambiguous boundaries, with adjustments for variation in complex or atypical anatomy and image artifacts. But this flexibility introduces variability in any manual protocol. Tracing both hippocampi can also take an expert several hours, resulting in prohibitive expense for today's large and growing MRI data sets. See (Turner, 2014) for a review on the growing size of imaging studies. In pursuit of a reliable technique that will scale up with growing studies, several algorithms have been created to automate the segmentation process. FreeSurfer was selected for its availability, ease of use, nearly full automation, and higher accuracy than alternative algorithms (Morey et al., 2009; Sánchez-Benavides et al., 2010; Schoemaker et al., 2016; Tae, Kim, Lee, Nam, & Kim, 2008; Wenger et al., 2014).

Automated techniques are often published with a brief validation of their output, but insufficient independent validations comparing manual tracing to automated replacements exist, particularly with evolving software. Existing studies of FreeSurfer's reliability use relatively small numbers of participants and older software versions (Morey et al., 2009; Sánchez-Benavides et al., 2010; Schoemaker et al., 2016; Tae et al., 2008; Wenger et al., 2014). We are unaware of any studies that assess validity of current FreeSurfer 6.0.0 software, hippocampal subfield algorithms, or assessment of handedness. To fully characterize the differences between modern FreeSurfer segmentations, prior FreeSurfer versions, and our manual tracing protocol, we compared manually traced left and right hippocampi to FreeSurfer 5.3.0 and 6.0.0 segmentations of MRIs from 664 participants from the Genetics of Microangiopathic Brain Injury (GMBI) study.

To investigate methodological similarities and differences directly, we compared raw and total intracranial volume (TIV)-adjusted hippocampal volumes by both value and voxel-wise percent agreement. To answer the more salient question of whether these methods' quantitatively different results provide accurate and comparable estimates for biologically relevant measures, we also investigated right-left asymmetry and an assessment of atrophy over the adult lifespan using data from each method. Finally, to determine the ongoing existence of previously reported age and volume biases (Wenger et al., 2014), we investigated consistency in differences between methods relative to differences in participants' ages and hippocampal volumes.

2 | METHODS

2.1 | Participants

The initial sample comprised 708 non-Hispanic white participants, age 37–93, 41% male, 93% right-handed, from Rochester, MN who underwent MRI between August 2001 and February 2006 as part of the GMBI study. Participants were recruited into GMBI from the Rochester Epidemiology Project (Melton, 1996), based on at least two members of each sibship having hypertension before age 60. The sample undergoing full analyses in this study was a subset of 664 images, age 38–84, 39% male, 94% right-handed. Exclusions are shown in Table 1. This study was approved by the Institutional Review Board of The Mayo Clinic, Rochester and The University of Mississippi Medical Center. All participants gave written informed consent prior to participation.

TABLE 1 Sex and age characteristics of the initial and final sample

	Male		Female		Total	
	n	age	*n*	age	*n*	age
Initial sample	288	62.7 (9.1)	420	61.2 (9.3)	708	61.8 (9.2)
Orientation issues	20	66.2 (6.3)	6	74.4 (8.2)	26	68.1 (7.5)
Quality issues	2	58.8	1	53.3	3	57.0
Tumor present			1	58.8	1	58.8
FreeSurfer errors	23	66.0	11	69.2	34	67.0
Missing TIV	11	64.8	7	71.1	18	67.2
Final sample	262	62.4 (9.2)	402	60.9 (9.3)	664	61.5 (9.2)

Ages are reported as arithmetic mean (*SD*). Exclusions sum to more than the total because some participants were excluded for multiple reasons.

2.2 | MRI acquisition

MRIs were captured with the same clinical GE Signa 1.5T MRI scanner, maintained over time with recommended software updates, as a component of the GMBI Study between 2001 and 2006. The complete protocol included several sequences as described previously (Knopman et al., 2008). This study used 0.9375 × 0.9375 × 1.6 mm voxel T1-weighted coronal 3D spoiled gradient echo (SPGR) images (256 × 192 × 124 matrix, 6–10 ms echo time, 24 ms repetition time, 25-degree flip angle, 24 × 18 × 19.8 cm field of view). TIV were previously measured manually from 2D T1-weighted sagittal images (32 slices, 5 mm thick; no gap, 256 × 192 matrix, 14–20 ms echo time, 500 ms repetition time, 24 cm field of view; Knopman et al., 2008). To rule out the potential for bias introduced by changing scanner software, we assessed hippocampal volume, TIV, contrast to noise ratio (CNR), and signal to noise ratio (SNR) over the image acquisition dates. All four relationships, after including age in the model, were insignificant, failing to support any effect due to scanner changes. Moreover, FreeSurfer's scan-rescan reliability has been shown to remain high even across significant hardware and software upgrades (Jovicich et al., 2009).

2.3 | Manual tracing

708 MRIs were initially available for analysis. We followed a previously published protocol for tracing hippocampal boundaries (Jack, Theodore, Cook, & McCarthy, 1995). Briefly, we used MIPAV (Medical Image Processing, Analysis, and Visualization) software (RRID: SCR_007371; McAuliffe et al., 2001) to orient the image coronally, then rotated the image around the left–right axis until the hippocampus aligned horizontally. We saved the reoriented image and traced the outline of each cross-sectional slice of hippocampus from the caudal end, where the

fornix separates from the hippocampus in the lateral ventricle, to the rostral end, where the hippocampal head diminishes beneath the amygdala. We saved the coordinates of all boundary points to calculate the area of each slice, then multiplied areas by the slice gap to approximate hippocampal volume.

MRIs were obtained in Analyze format. For 26 (3.7%) of 708 images, neurologic versus radiologic orientation could not be determined reliably. They were excluded along with three images with extensive artifacts that could not be traced, one image with a large tumor, and 5 (0.7%) images with no TIV measurements available, leaving 673 manually-traced images (Table 1).

We distributed the workload of manually tracing hippocampi between two trained experts, MFS and KBF, both authors of this manuscript. A third tracer, BD, was also trained, but contributed tracings only for inter-rater reliability, including Figure 1, in this study. These tracings were periodically recalibrated by reviewing and discussing segmentations. Repeated intra-rater and inter-rater reliability studies found consistent agreement between scorers and no sign of drift over time. TIV were measured previously (Smith et al., 2009; Turner et al., 2005).

2.4 | Automated segmentation

All 708 MRIs available for manual tracing, including those known to be problematic, were segmented by FreeSurfer versions 5.2.0, 5.3.0, and 6.0.0 (RRID: SCR__01847; Fischl et al., 2002). Each 3D SPGR image was processed with default settings on the Mississippi Center for Supercomputing Research SuSe Linux cluster, but with additional flags for supplemental hippocampal-subfield and brainstem-structures routines in version 6.0.0. Each recon-all process was allocated a single Xeon E5420 2.5 GHz processor and 6GB RAM. FreeSurfer is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu) and has been thoroughly described elsewhere (Fischl et al., 2002; Fischl, 2012, Iglesias et al., 2015). Scripts used to manage this project are also freely available online (Schmidt, 2017).

FreeSurfer 6.0.0 failed with errors on 25 (3.5%) images. FreeSurfer 5.3.0 failed on the same 25, and one additional. FreeSurfer 5.2.0 failed on 32 (4.5%), including all images with errors on 5.3.0. Two (0.3%) images, successfully processed with FreeSurfer 5.2.0 and 6.0.0, generated no errors with FreeSurfer 5.3.0, but failed to complete on multiple attempts. Error-free output from FreeSurfer was manually reviewed by displaying each segmentation result (mri/aseg.mgz file) in tkmedit, part of the FreeSurfer software suite. This quality control step revealed 6 (0.8%) images with incorrect orientation, and two (0.3%) with insufficient quality or contrast. Four (0.6%) images had severe ventriculomegaly, but were not excluded as the segmentation was successful by all methods. FreeSurfer 6.0.0 data were used to represent FreeSurfer in all comparisons to manual tracing, unless specified otherwise. Standard hippocampal volumes were extracted from the "Left-Hippocampus" and "Right-Hippocampus" fields in FreeSurfer's aseg.stats file for each of three software versions.

We also compared two volumes assembled from FreeSurfer 6.0.0's hippocampal subfields processing stream (Iglesias et al., 2015). The

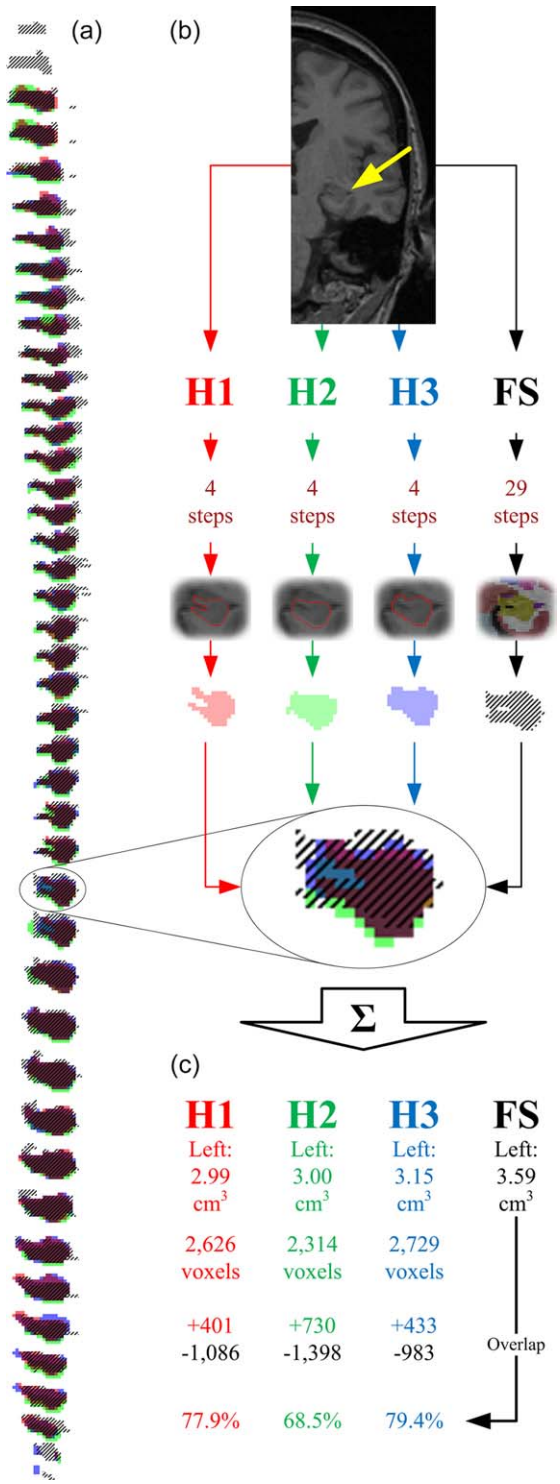

FIGURE 1 Method of comparing overlap. (a) All slices of one left hippocampus from the tail (posterior) on top to the head (anterior) on bottom, as traced by three humans (colors) and segmented by FreeSurfer (hatched). (b) One representative hippocampal slice, circled in a, is expanded into each scorer's assessment. (c) Volumes are calculated by summing all slices for each scorer, reported with overlapping voxel counts, then voxels by human and FreeSurfer individually. Overlap is reported as Dice's coefficient. H1, H2, and H3 indicate three independent human scorers; FS indicates FreeSurfer. All slices of this representative hippocampus are shown enlarged in Supporting Information Figure S1

"subfields" or "sf+" volumes come from FreeSurfer's "Whole_hippocampus" field in the hippoSfVolumes files. As the most obvious difference between FreeSurfer's atlas and our protocol's hippocampal boundaries is FreeSurfer's inclusion of tail tissue, we derived a "subfields-T" or "sf- "volume, by subtracting the "Hippocampal_tail" field from the "Whole_hippocampus" field. It is intended to represent a closer approximation, and fairer comparison, to our manual traces.

2.5 | Comparisons

Comparisons between automated and manual methods were conducted four ways, all in R (RRID:SCR_001905; R Core Team, 2017). First, direct comparisons were made between raw volumes reported by each method. Multiple comparison algorithms were employed due to differing strengths and weaknesses of each. (1) Pearson pairwise correlations (r) and (2) Simple linear regression (r^2) measure the tendency for one variable to rise or fall with another, making them good measures of agreement for studies more concerned with change in volume than raw volume. (3) paired t tests (p) demonstrate whether means differ between raters. (4) Intraclass correlation coefficients (ICC) were calculated using the R "psych" package (Revelle, 2017), as defined by Shrout and Fleiss for a two-way random effects model, also called ICC(2,1) (Shrout & Fleiss, 1979). ICCs are considered the best tool for inter-rater comparisons when comparable raters provide comparable values. FreeSurfer and manual tracing, however, have historically provided significantly different values. This between-group difference in means is penalized by ICC, resulting in lower ICC scores even when volumes are well-correlated.

Raw hippocampal volumes are reported in all analyses. TIV-adjusted hippocampal volumes (Equation 1) are added where appropriate. TIV-adjustments are calculated by regression rather than simple volume/TIV ratios as described previously (Jack et al., 1989; Arndt, Cohen, Alliger, Swayze, & Andreasen, 1991). Adjusting based on regression assures the exact adjustment needed to cancel out the relationship to TIV, is more likely to result in normally distributed adjusted volumes, and maintains the units and approximate range of the original volumes. This calculation required first regressing hippocampal volumes on TIVs to obtain a β coefficient, then adjusting hippocampal volumes on the difference between each individual's measured TIV and its expected value relative to measured hippocampal volume. Manual hippocampal volumes were adjusted to manual TIVs; FreeSurfer volumes were adjusted to Freesurfer-supplied eTIVs; right and left hippocampal volumes were adjusted independently.

$$\mathrm{HVol}_{\mathrm{adj}} = \mathrm{HVol}_{\mathrm{raw}} - \beta(\mathrm{TIV}_{\mathrm{raw}} - \mathrm{TIV}_{\mathrm{mean}}) \quad (1)$$

$\mathrm{HVol}_{\mathrm{raw}}$ and $\mathrm{TIV}_{\mathrm{raw}}$ refer to the measured values for a given participant. $\mathrm{TIV}_{\mathrm{mean}}$ refers to the mean TIV of the sample.

Second, to attain a more accurate assessment of agreement, respecting location in addition to volume, overlap of specific voxels was assessed. Dice's coefficient is the only comparison algorithm in this study able to account for agreement on the location of the hippocampus within the MRI. To generate Dice's coefficients, manually traced images were first registered onto the FreeSurfer aseg.mgz with AFNI tools (Cox, 1996; Cox & Hyde, 1997; Gold et al., 1998) to ensure coordinate consistency. Masks were created from the manual tracings using MIPAV and compared to an uncompressed copy of FreeSurfer's labeled aseg.mgz image in three dimensions with Gnu Octave (Eaton, Bateman, Hauberg, & Wehbring, 2015). Scripts used for this process are freely available online (Schmidt, 2017). The ratio below is expressed as Dice's coefficient (Dice, 1945; Equation 2) as performed in the original FreeSurfer report (Fischl et al., 2002; Figure 1)

$$\mathrm{Dice} = \frac{|A \cap B| + |B \cap A|}{|A| + |B|} \quad (2)$$

A represents voxels selected by segmentation A; B represents voxels selected by segmentation B. $|A \cap B|$ represents the voxels selected by both A and B.

Third, an asymmetry index was calculated as suggested by prior literature (Hasan & Pedraza, 2009; Equation 3) and an additional Cohen's d value (Equation 4) was calculated for comparisons to published meta-analyses (Pedraza, Bowers, & Gilmore, 2004). Statistical significance is reported as a one-sample t test with a null hypothesis of asymmetry = 0.

$$\mathrm{asymmetry} = \frac{2(\mathrm{vol}_{\mathrm{right}} - \mathrm{vol}_{\mathrm{left}})}{(\mathrm{vol}_{\mathrm{right}} + \mathrm{vol}_{\mathrm{left}})} \quad (3)$$

$$d = \left(1 - \frac{3}{4n-9}\right)\left(\frac{\mathrm{vol}_{\mathrm{right}} - \mathrm{vol}_{\mathrm{left}}}{\sigma_{\mathrm{pooled}}}\right) \quad (4)$$

Fourth, we assessed volume by each technique over age, and compared the level of atrophy calculated from each method. Change cannot be directly measured in our cross-sectional sample, so annualized atrophy was quantified with a generalized linear model regressing volume on age with a log link function and a gamma distribution. Exponentiated coefficients represent percentage annualized difference in volume, or "atrophy." A value of 0.010 implies that 1.0% of volume is being lost each year.

To assess stability and consistency of segmentation algorithms over age and volume, data from each method of segmentation, and comparison measures between methods, were compared to age and traced volumes via linear regression models. Comparison measures and log-linear relationships were also compared across age strata: <55; 55–70; >70 as assessed in a recent meta-analysis (Fraser, Shaw, & Cherbuin, 2015).

Power analyses were performed in R with the "pwr" package.

3 | RESULTS

Of the 708 MRIs available for analysis, 664 remained after all exclusions. Our sample was 39% male, averaged 61 years of age (SD = 9.2 years), and was exclusively non-Hispanic White from Rochester, MN. The volumes reported from each method are in Table 2 (Figure 2). Manual tracing resulted in significantly smaller (left Δ −641 mm^3, 95%CI: −663 to −618, $p < .0001$; right Δ −680 mm^3, 95%CI: −703 to −657, $p < .0001$) volumes than FreeSurfer 6.0.0. We detected FreeSurfer biases relative to age and size (Figure 3). The difference between FreeSurfer and tracing diminished with increasing

TABLE 2 Hippocampal and total intracranial volumes measured by different methods

	R Hippo	L Hippo	TIV
Manual	3,293 (*SD* = 424) mm^3	3,210 (*SD* = 397) mm^3	1,454,018 (*SD* = 144,086) mm^3
FS 6.0.0	3,973 (*SD* = 433) mm^3	3,851 (*SD* = 405) mm^3	1,571,552 (*SD* = 199,205) mm^3
FS 6 sf+	3,465 (*SD* = 382) mm^3	3,363 (*SD* = 363) mm^3	
FS 6 sf-	2,911 (*SD* = 327) mm^3	2,838 (*SD* = 314) mm^3	
FS 5.3.0	4,056 (*SD* = 478) mm^3	3,995 (*SD* = 472) mm^3	1,561,957 (*SD* = 205,156) mm^3
FS 5.2.0	4,046 (*SD* = 468) mm^3	3,986 (*SD* = 469) mm^3	1,561,943 (*SD* = 205,158) mm^3

sf+ indicates sum of all hippocampal subfields; sf− is the same, with the tail subtracted. See Figure 2.

traced volume in versions 5.3.0 (−0.15 mm^3 difference per mm^3 volume increase, 95%CI: −0.20 to −0.09, $p < .001$) and 6.0.0 (−0.21 mm^3 difference per mm^3 increase, 95%CI: −0.26 to −0.17, $p < .001$). The difference also decreased with increasing age in FreeSurfer version 6.0.0 (−12.1 mm^3 smaller difference per year of age, 95%CI: −16.5 to −7.8, $p < .0001$), although to a lesser extent than in version 5.3.0 (−22.8 mm^3 smaller difference per year of age, 95%CI: −27.5 to −18.0, $p < .0001$). The age bias is strengthened by including a TIV adjustment in the linear model (6.0.0: −17.3 mm^3, smaller difference per year of age, 95%CI: −21.4 to −13.2, $p < .0001$; 5.3.0: −27.1 mm^3, 95%CI: −31.7 to −22.5, $p < .0001$).

Surprisingly, the mean "whole_hippocampus" measure from FreeSurfer 6.0.0's hippocampal subfields routine (sf+) was also significantly smaller than FreeSurfer's default "aseg" values (left Δ −488 mm^3, 95%CI: −500 to −476, $p < .0001$; right Δ −508 mm^3, 95%CI: −520 to −496, $p < .0001$) and only slightly larger than traced values (left Δ 153 mm^3, 95%CI: 133–173, $p < .0001$; right Δ 172 mm^3, 95%CI: 152–192, $p < .0001$; Figure 2). Although sf+ measures were nearer traced volumes in value, their biases by volume (−0.27 mm^3 difference per mm^3 increase, 95%CI: −0.32 to −0.23, $p < .0001$) and age (−12.81 mm^3 smaller difference per year of age, 95%CI: −16.7 to −8.9, $p < .0001$) were similar to aseg measures (Figure 4).

Pearson's correlations between manually traced hippocampi and FreeSurfer 6.0.0-segmented hippocampi were $r = .77$ (left $r = .73$, right $r = .76$). ICCs were 0.32 on the left and 0.34 on the right, kept low by the between group volume difference. And linear regression reported $r^2 = .60$ (left $r^2 = .54$, right $r^2 = .57$; Table 3). All three measures were higher for FreeSurfer 6.0.0 than 5.3.0, but even higher for sf+. Hippocampal volumes from FreeSurfer 6.0.0 are also smaller than those from FreeSurfer 5.3.0 (Left Δ −144.0 mm^3, 95%CI: −159.9 to −128.2, $p < .0001$; Right Δ −83.3 mm^3, 95%CI: −98.3 to −68.4, $p < .0001$), which were comparable (Left Δ 8.7 mm^3, 95%CI: −3.6 to 21.0, $p = .165$; Right Δ 9.9 mm^3, 95%CI: −1.7 to 21.5, $p = .095$) to those from version 5.2.0.

Rightward asymmetry was detected in both right and left-handed subgroups by both manual tracing (right-handed: $n = 620$, +2.51%, 95%CI: 2.02%–3.01%, $p < .0001$; left-handed: $n = 40$, +1.70%, 95%CI: −1.17% to 4.57%, $p = .238$) and FreeSurfer 6.0.0 (right-handed: +3.08%, 95%CI: 2.63%–3.53%, $p < .0001$; left-handed: +2.79%, 95%CI: 0.46%–5.12%, $p = .020$). But smaller numbers of left-handed participants resulted in reduced confidence and loss of statistical significance in manually traced left-handed participants. (Table 4). In the entire sample, right hippocampi were larger than left in all methods ($p < .001$; Figure 2).

FreeSurfer volumes overlapped traced volumes with Dice's coefficient of 76% between right hippocampi and 75% between the left in both versions 6.0.0 ($n = 700$ comparisons) and 5.3.0 ($n = 699$). These agreements remained the same for right-handed and left-handed participants. Overlap between 6.0.0 and 5.3.0 was 88% on the left and 87% on the right. FreeSurfer 6.0.0's sf+ measure showed slightly higher overlap with manual traces of 78% on both right and left, and overlapped FreeSurfer 6.0.0's aseg volume at 83% on the left and 84% on the right. We assessed inter-rater reliability by comparing overlap for 13 images traced multiple times by different tracers ($n = 25$ comparisons), resulting in a Dice's coefficient of 82% on both right and left sides. Intra-rater reliability came from repeated tracings of four images, all by the same author ($n = 22$ comparisons) with Dice's coefficients of 88% on the right, and 86% on the left. FreeSurfer is a deterministic process, reporting identical volumes on each assessment of the same input image, resulting in intra-rater agreement of 100%. All

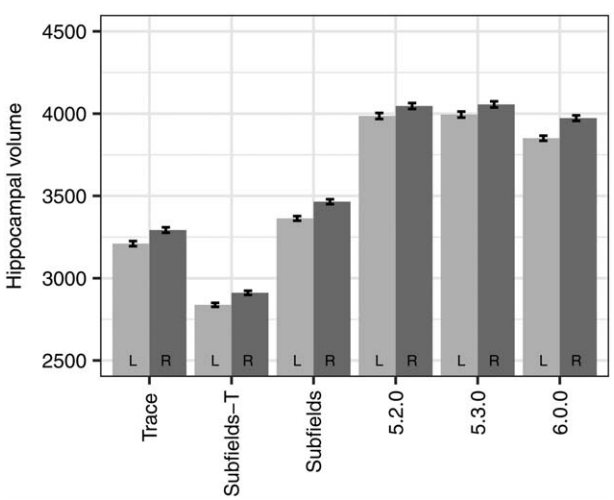

FIGURE 2 Different methods of segmentation result in significantly different values for raw hippocampal volume. All versions of FreeSurfer report larger volumes than manual tracing. FreeSurfer's hippocampal subfield processing is closer to manual tracings than its standard subcortical segmentation stream. Right hippocampi are also slightly larger according to all methods. Error bars represent standard error of the mean

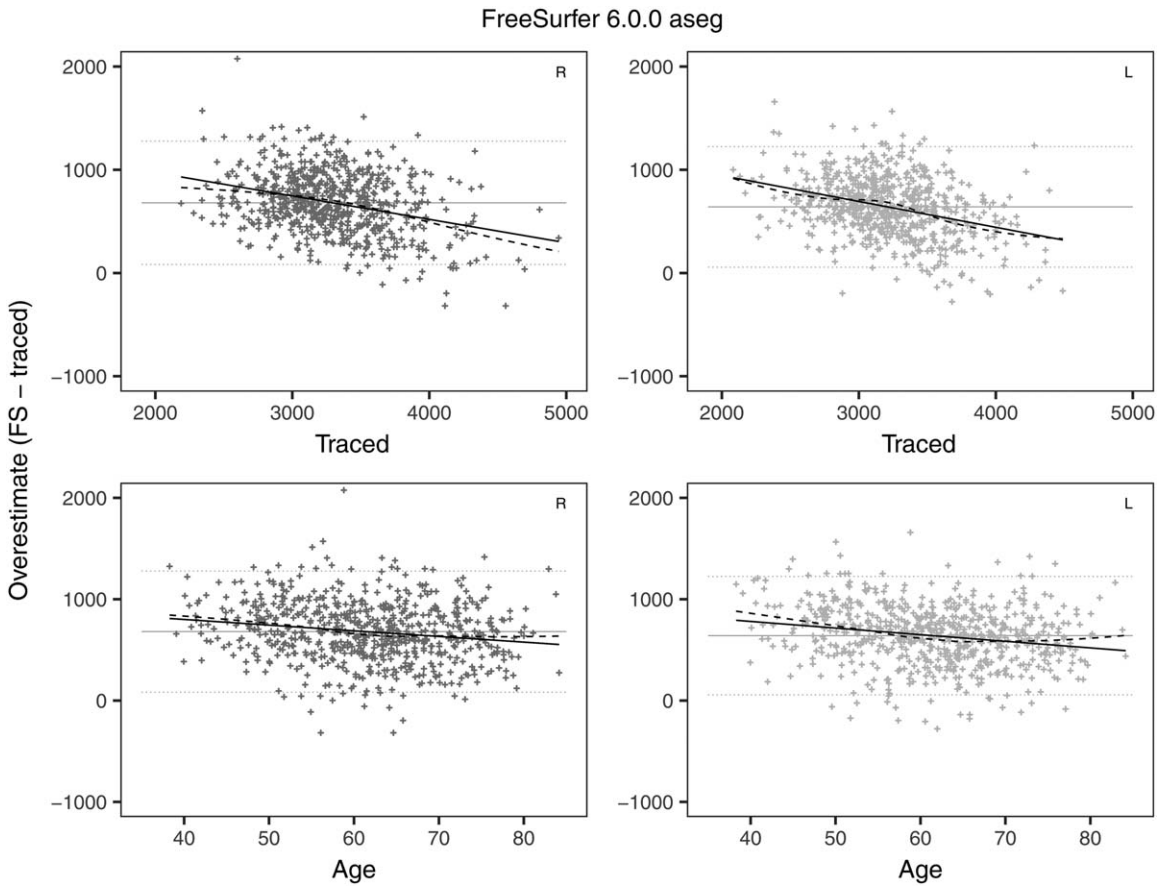

FIGURE 3 FreeSurfer's overestimate differs with volume and age. Bland-Altman plots show the difference between raw volumes from FreeSurfer 6.0.0's subcortical segmentation stream and manual tracing on the *y*-axis. That difference is reduced in larger hippocampi (top) or increasing age (bottom) in both hemispheres

comparisons are shown in Table 3 and an illustration of overlap is shown in Figure 1. See Supporting Information Figure S1 for all slices.

Estimates of hippocampal atrophy differed only slightly between segmentation methods. Manual tracing reports an overall 0.31% (95% CI: 0.21%, 0.41%) annualized atrophy rate (Left 0.28% [95% CI 0.18%, 0.38%], Right 0.34% [95% CI: 0.24%, 0.44%]) compared to 0.41% (95% CI: 0.33%, 0.49% $p = .117$) annualized atrophy (Left 0.40% [95% CI: 0.32%, 0.49% $p = .066$], Right 0.42% [95% CI: 0.34%, 0.51% $p = .230$]) via FreeSurfer 6.0.0. Assessed separately by age group, atrophy estimates are higher in older participants and lower in younger participants, regardless of segmentation method. (Figure 5 and Table 5)

4 | DISCUSSION

FreeSurfer segmentations generated larger hippocampal volumes and masks than those from our manual tracing protocol. Interestingly, using volumes from the hippocampal subfield protocol reduced that discrepancy and increased agreement with manual tracings. Volumes from each method were well-correlated, however, with adequate overlap, and biases detected in earlier versions of FreeSurfer have improved slightly in version 6.0.0. Most importantly, all techniques reported smaller hippocampal volumes and higher atrophy rates in older versus younger participants, and rightward asymmetry. Within the demographics represented by our cohort, these results suggest that, while quantitatively different, FreeSurfer and manual tracing support similar answers to **the** biologically relevant estimation of atrophy.

Our results agree with previous literature that FreeSurfer reports larger volumes than manual protocols (Cherbuin, Anstey, R?Glade-Meslin, & Sachdev, 2009; Morey et al., 2009; Sánchez-Benavides et al., 2010; Tae et al., 2008; Wenger et al., 2014) and that right hippocampi are slightly larger than left, on average (Hasan & Pedraza, 2009; Jack et al., 1989; Li, Ga, Huo, Li, & Gao, 2007; Tae, Kim, Lee, Nam, & Kim, 2009). Our estimates of annualized atrophy, 0.31% with manual tracing and 0.41% with FreeSurfer both fit into the low end of previous reports, which range from 0.28% to 6.38% per annum (Fraser et al., 2015; Raz et al., 2004a; Raz & Rodrigue, 2006). Atrophy rates in AD patients, for comparison, have been reported at 3%–8% per annum (Barnes et al., 2009; Raz & Rodrigue, 2006). Our low atrophy could be explained by our sample's mean age of 61.3, about a decade younger than most studies of aging and dementia. Stratifying by age reveals higher atrophy in older members of our cohort, 1.1% per annum in 129 participants over 70, similar to the 1.1% reported in controls over 70 (Fraser et al., 2015) and 1.4% reported for similarly-aged controls in AD studies (Barnes et al., 2009; Figure 6). Stratification suggests that the association between hippocampal volume and age is nonlinear

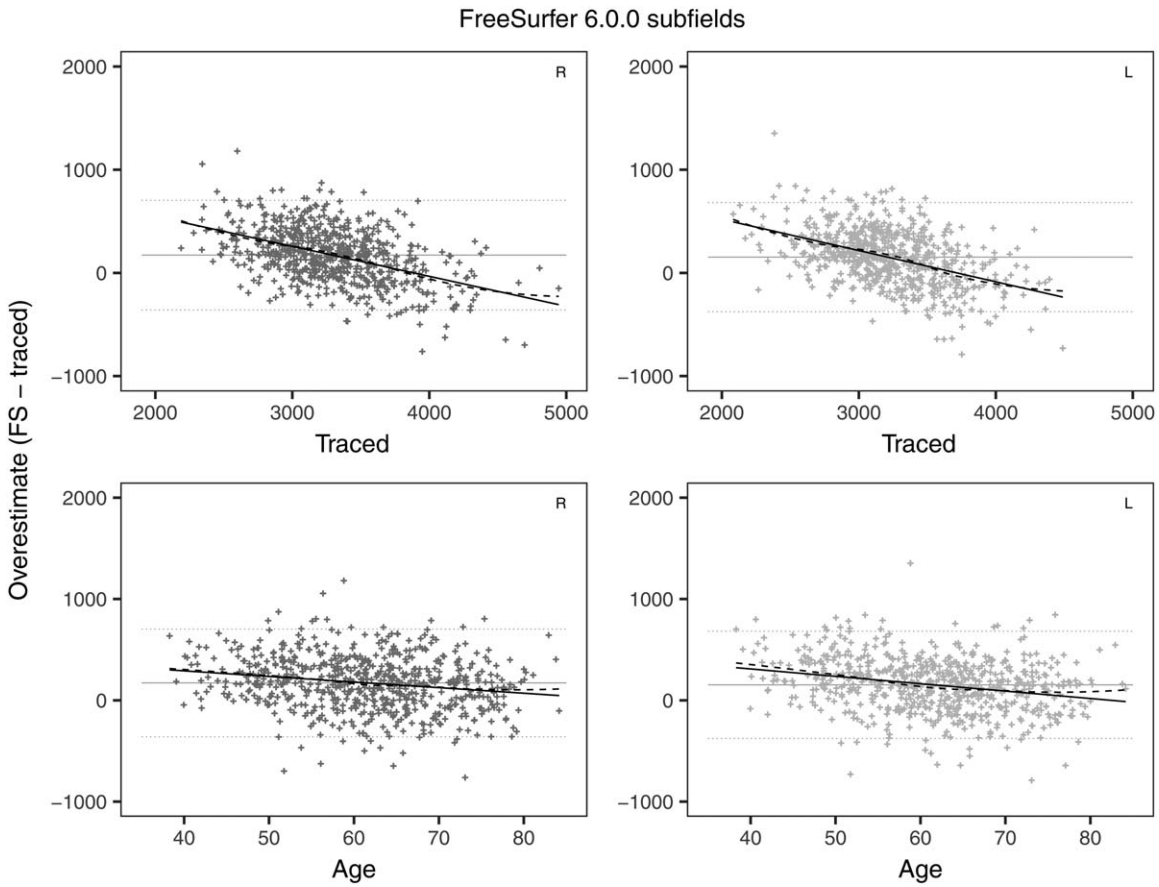

FIGURE 4 FreeSurfer's hippocampal subfield overestimates are smaller, but also differ with volume and age. Bland-Altman plots show the difference between raw volumes from FreeSurfer 6.0.0's hippocampal subfields and manual tracing on the *y*-axis. Although there is little difference between subfields and manual tracing, on average, the difference is reduced over size and age similarly to FreeSurfer's subcortical segmentation stream

(Figure 5), in agreement with (Walhovd et al., 2005), who used FreeSurfer segmentation, and (Raz et al., 2004a), who manually traced. However, both our manual and automated results suggest an opposite relationship prior to age 40 than that suggested by (Allen et al., 2005), who used manual tracing. All comparison studies reported on healthy participants across the adult lifespan.

FreeSurfer was selected from several automated segmentation tools. It is widely and freely available, well-automated, well-documented, actively developed, and easy to implement. Previous studies reported that earlier versions of FreeSurfer agree more closely with manual methods than other common tools, FSL/FIRST (Morey et al., 2009; Schoemaker et al., 2016) or IBASPM (Tae et al., 2008). Some hippocampus-specific classification systems (Tangaro et al., 2014; Platero & Tobar, 2016) and the automatic brain segmentation system (Hosseini et al., 2016) have outperformed FreeSurfer in hippocampus-specific segmentations. FreeSurfer has also been shown to be more sensitive to the atrophy coincident with MDD than alternatives (Morey et al., 2009). Independent scans of the same individual show imperfect, but consistent agreement between FreeSurfer analyses (Morey et al., 2010). These advantages combined with the benefit of continuing active development made FreeSurfer our best option for automated hippocampal segmentation. The most recent version, 6.0.0, reduces the volume discrepancy (Figure 2) and volume bias (Figure 3). There are no perfect tools, however. Investigators are responsible for basic manual validation of any tool's output and FreeSurfer provides no exception. FreeSurfer documentation (available online) and training prescribe manually checking several steps throughout its process to ensure reliable segmentations. In this study, 708 images were submitted to FreeSurfer, including 27 images with incorrect orientation data. 21 of the 27 misoriented images generated errors and failed, making manual checks unnecessary, but six images were registered sideways to the atlas, completed the process "successfully" and generated incorrect data. These were only excluded from our analyses after manual review of all segmentations. Two other images caused FreeSurfer to fail, but generated no error messages, leaving partial data for analysis and reinforcing the need for manual review. The necessary step of manually quality-checking output requires under a minute per observation, comparing favorably to the hours required for manual tracing. Researchers reap many benefits from automation, but expecting complete autonomy from any software is irresponsible and increases the risk of noisy data and misleading results. These potential issues do not represent a problem with FreeSurfer, but reinforce the need for users to manually review, and even edit intermediates, before using final FreeSurfer measurements to test hypotheses.

TABLE 3 Comparisons of agreement between methods

		r		ICC		Dice	
		R	*L*	*R*	*L*	*R*	*L*
Trace comparisons							
Intra-rater	(n = 22)	0.69	0.90	0.84	0.95	0.88 (0.025)	0.86 (0.025)
Inter-rater	(n = 25)	0.68	0.69	0.59	0.59	0.82 (0.047)	0.82 (0.048)
Trace	6.0.0	0.73	0.72	0.34	0.32	0.76 (0.042)	0.75 (0.044)
Trace	sf+	0.78	0.77	0.72	0.70	0.78 (0.042)	0.78 (0.044)
Trace	sf–	0.78	0.77	0.51	0.48	0.75 (0.041)	0.76 (0.044)
Trace	5.3.0	0.71	0.70	0.29	0.27	0.76 (0.046)	0.75 (0.052)
Trace	5.2.0	0.72	0.70	0.31	0.27	0.76 (0.044)	0.75 (0.052)
FreeSurfer 6.0.0 comparisons							
6.0.0	sf+	0.92	0.89	0.52	0.51	0.84 (0.036)	0.83 (0.037)
6.0.0	sf–	0.91	0.89	0.19	0.18	0.77 (0.031)	0.77 (0.032)
6.0.0	5.3.0	0.87	0.89	0.89	0.84	0.87 (0.076)	0.86 (0.082)
6.0.0	5.2.0	0.90	0.89	0.92	0.86	0.88 (0.061)	0.87 (0.067)

Comparisons between multiple tracings are divided into intra-rater (the same human rater re-tracing the same hippocampus) and inter-rater (different humans tracing the same hippocampus). Manual tracings and FreeSurfer version 6.0.0 were compared to all other methods. Dice's coefficients are reported with standard deviations in parentheses. Older and alternate measurement comparisons are excluded. sf+ refers to hippocampal subfields' "whole_hippocampus" field; sf– refers to sf+ with tail subtracted. See methods for caveats regarding interpretation.

Manual tracing, broadly defined, is considered the gold standard in segmentation studies due largely to the human ability to deal appropriately with abnormal data. But there are at least 71 published protocols for manually tracing hippocampi (Boccardi et al., 2011; Konrad et al., 2009), and even the same protocol can result in different volumes if performed at different angles through the image (Hasboun et al., 1996). Defining any of these protocols, or none, as the narrowly defined gold standard is controversial (Boccardi et al., 2011; Boccardi et al., 2015b; Hasan, 2009) as objective reality depends on varying definitions of anatomical boundaries in addition to the technical anomalies inherent to any protocol. Despite this controversy, the most commonly cited protocols were merged into a single harmonized protocol by the European Alzheimer's Disease Consortium (EADC) and Alzheimer's Disease Neuroimaging Initiative (ADNI) between 2008 and 2013 (Boccardi et al., 2011; Boccardi et al., 2015a; Boccardi et al., 2015b; Frisoni et al., 2015; Frisoni & Jack, 2011). Our tracing work began in 2012 and employed a protocol (Boccardi et al., 2011; Jack, 1994) that was later merged into the EADC-ADNI harmonized protocol. It is precisely defined and well-established, but it differs from the Center for Morphometric Analysis (CMA) guidelines (Makris et al.,) underlying FreeSurfer's atlas. First, (Jack, 1994) and (Watson et al., 1992) define the caudal boundary of the hippocampus as "the coronal section in which the crus of the fornix is seen in full profile" while CMA recommends inclusion of further caudal regions, resulting in larger tails. Second, FreeSurfer segmentations tend to include partial voxels between hippocampus and lateral ventricle that are more often excluded in manual tracings (Cherbuin et al., 2009; Han & Fischl, 2007; Tae et al., 2008). Both of these differences contribute to FreeSurfer's larger volumes, which were 20% larger than their manual counterparts using 6.0.0, and 24% larger using 5.3.0, consistent with previous reports of 26% with FreeSurfer 4.3 (Cherbuin et al., 2009) and 38% with FreeSurfer 3.04 (Tae et al., 2008).

Regardless of manual tracing's status as the gold standard, the labor cost of hand-tracing is prohibitive for large studies, making automation the default method, and periodic re-calibration studies like this one critical. Manual hippocampal tracing costs approximately $23.00 per subject (2 hippocampi × 1 hr each × $23,376 per year/2,000 paid hours per year), based on recommended graduate student stipend rates

TABLE 4 Asymmetry by method

	Difference (Δ)	Relative (%)		
	All	All	Right-handed (n = 620)	Left-handed (n = 40)
Manual	83 mm^3 (95% CI:67–98)	2.5% (2.0, 3.0%)***	2.5% (2.0, 3.0) d = 0.20 ***	1.7% (–1.2, 4.6) d = 0.14
FS 6.0	122 mm^3 (95% CI:105–139)	3.1% (2.6, 3.5%)***	3.1% (2.6, 3.5) d = 0.29 ***	2.8% (0.5, 5.1) d = 0.31
FS 6 sf+	102 mm^3 (95% CI:89–115)	3.0% (2.6, 3.4%)***	2.9% (2.5, 3.3) d = 0.26 ***	3.6% (1.2, 6.0) d = 0.39 *
FS 6 sf-	73.0 mm^3 (95% CI:61–85)	2.5% (2.1, 2.9%)***	2.5% (2.0, 2.9) d = 0.22 ***	3.2% (0.8, 5.7) d = 0.35
FS 5.3	62 mm^3 (95% CI:40–84)	1.5% (1.0, 2.1%)***	1.6% (1.1, 2.2) d = 0.14 ***	0.2% (–3.0, 3.4) d = 0.03
FS 5.2	60 mm^3 (95% CI:40–81)	1.5% (1.0, 2.1%)***	1.6% (1.0, 2.1) d = 0.13 ***	1.0% (–1.4, 3.4) d = 0.10

Positive asymmetry values indicate right > left. The large number of right-handers results in $p < .0001$ for every method. Fewer left-handers results in only sf+ being significant $p < .01$. sf+ indicates the sum of hippocampal subfields; sf– is the same, with the tail subtracted. Cohen's d statistic for effect size is included for comparison to meta-analyses. *$p < .01$; **$p < .001$; ***$p < .0001$.

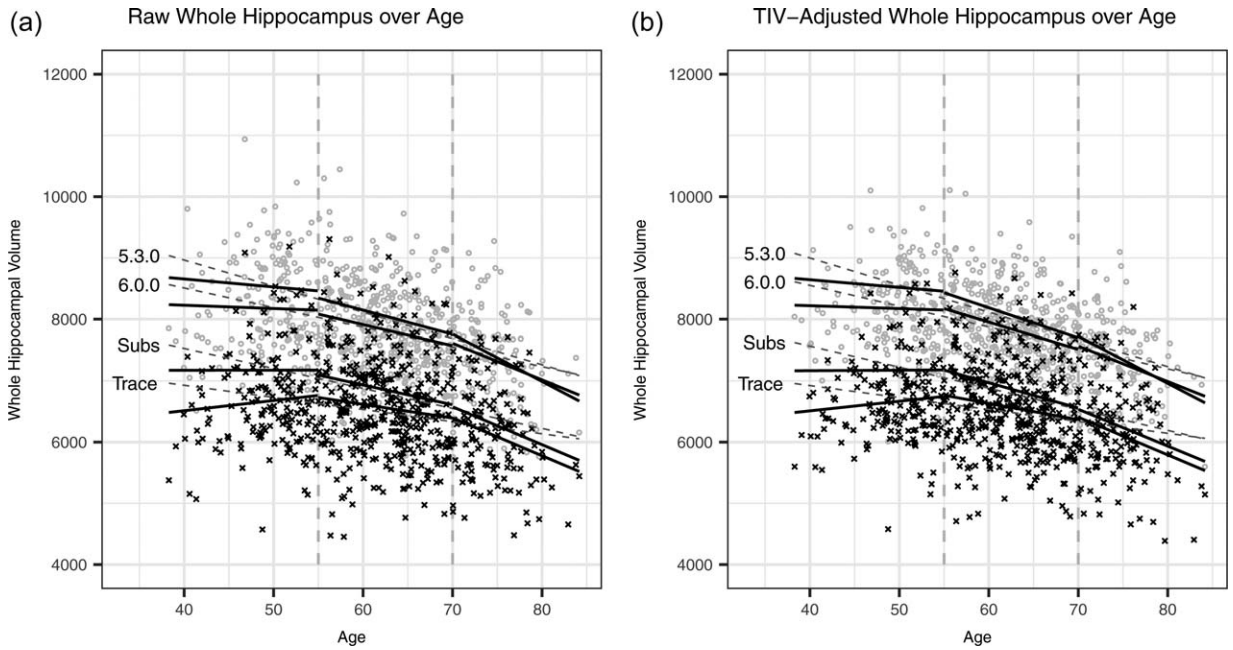

FIGURE 5 All methods report similar atrophy regardless of size differences. Hippocampal volumes are smaller in older participants than younger, on average, and the relationship between volume and age appears non-linear, steepening with increasing age. Gray dashed lines (labeled at their left edge) represent linear models pooled by method (Trace = Manually traced, Subs = Hippocampal Subfields, 6.0.0 = FreeSurfer 6.0.0, 5.3.0 = FreeSurfer 5.3.0); solid black lines represent linear models pooled for each age stratum (0–55, 55–70, and 70+) within each method. Gray O's represent individual FreeSurfer 6.0.0 volumes; black X's represent individual manually traced volumes; neither hippocampal subfields nor 5.3.0 volumes are plotted individually. In all three age groups, whether using raw values (a) or adjusted to TIV (b), FreeSurfer or manually traced, atrophy rates are shallower or even positive in younger subsamples and steeper in older subsamples. All methods and adjustments are qualitatively consistent regarding atrophy. See Table 5 for quantitative values

(NIH, 2016). Our FreeSurfer 6.0.0 processing, including 59.4 min (*SD* = 7.9) for hippocampal subfields, took an average of 20.2 hr per subject (SD = 4.4 hr). FreeSurfer 5.3.0 required 17.6 hr (*SD* = 2.7 hr) without hippocampal subfield processing. FreeSurfer 5.2.0 took only 8.4 hr (*SD* = 3.5 hr) per participant, but was run on faster hardware. Though our computing costs were donated, costs of running FreeSurfer on commodity computing resources, like Amazon's Elastic Compute Cloud would require approximately $1.89 (1 run on t2.large instance for 9.4 cents per hour × 20.2 hr) or $1.43 (15 concurrent runs on a r4.4 × large instance for $1.06 per hour × 20.2 hr) per subject (Amazon Inc., 2017) once the free software pipeline (Schmidt, 2017) has been set up. Moreover, FreeSurfer provides segmentations for dozens

TABLE 5 Stratified atrophy measures

	Raw hippocampal atrophy (per year)			
Strata	**Traced**	**FS 6.0.0**	**sf+**	**FS 5.3.0**
All (n = 664)	−19.81 mm^3 (−0.31%)	−31.93 mm^3 (−0.41%)	−32.62 mm^3 (−0.49%)	−42.57 mm^3 (−0.54%)
<55 (n = 171)	+16.65 mm^3 (+0.26%)	−5.37 mm^3 (−0.07%)	+0.18 mm^3 (0.00%)	−12.97 mm^3 (−0.16%)
55–70 (n = 364)	−22.42 mm^3 (−0.34%)	−34.05 mm^3 (−0.43%)	−32.38 mm^3 (−0.47%)	−39.27 mm^3 (−0.49%)
>70 (n = 129)	−61.84 mm^3 (−1.05%)	−59.68 mm^3 (−0.83%)	−61.83 mm^3 (−1.00%)	−76.85 mm^3 (−1.07%)
	TIV-adjusted hippocampal atrophy (per year)			
Strata	**Traced**	**FS 6.0.0**	**sf+**	**FS 5.3.0**
All (n = 664)	−19.49 mm^3 (−0.30%)	−33.91 mm^3 (−0.44%)	−34.37 mm^3 (−0.51%)	−44.19 mm^3 (−0.55%)
<55 (n = 171)	+15.73 mm^3 (+0.25%)	−4.68 mm^3 (−0.04%)	+0.79 mm^3 (+0.03%)	−12.54 mm^3 (−0.14%)
55–70 (n = 364)	−26.53 mm^3 (−0.39%)	−44.18 mm^3 (−0.55%)	−41.3 mm^3 (−0.59%)	−49.03 mm^3 (−0.60%)
>70 (n = 129)	−61.56 mm^3 (−1.05%)	−57.78 mm^3 (−0.81%)	−60.16 mm^3 (−0.98%)	−76.41 mm^3 (−1.07%)

Regression coefficients for each age stratum and method, expressed as mm^3, calculated by linear glm, and plotted in Figure 5. The same data are also expressed as a percentage of existing volume lost each year, calculated with a gamma family log-link glm, and plotted in Figure 6. glm = generalized linear model.

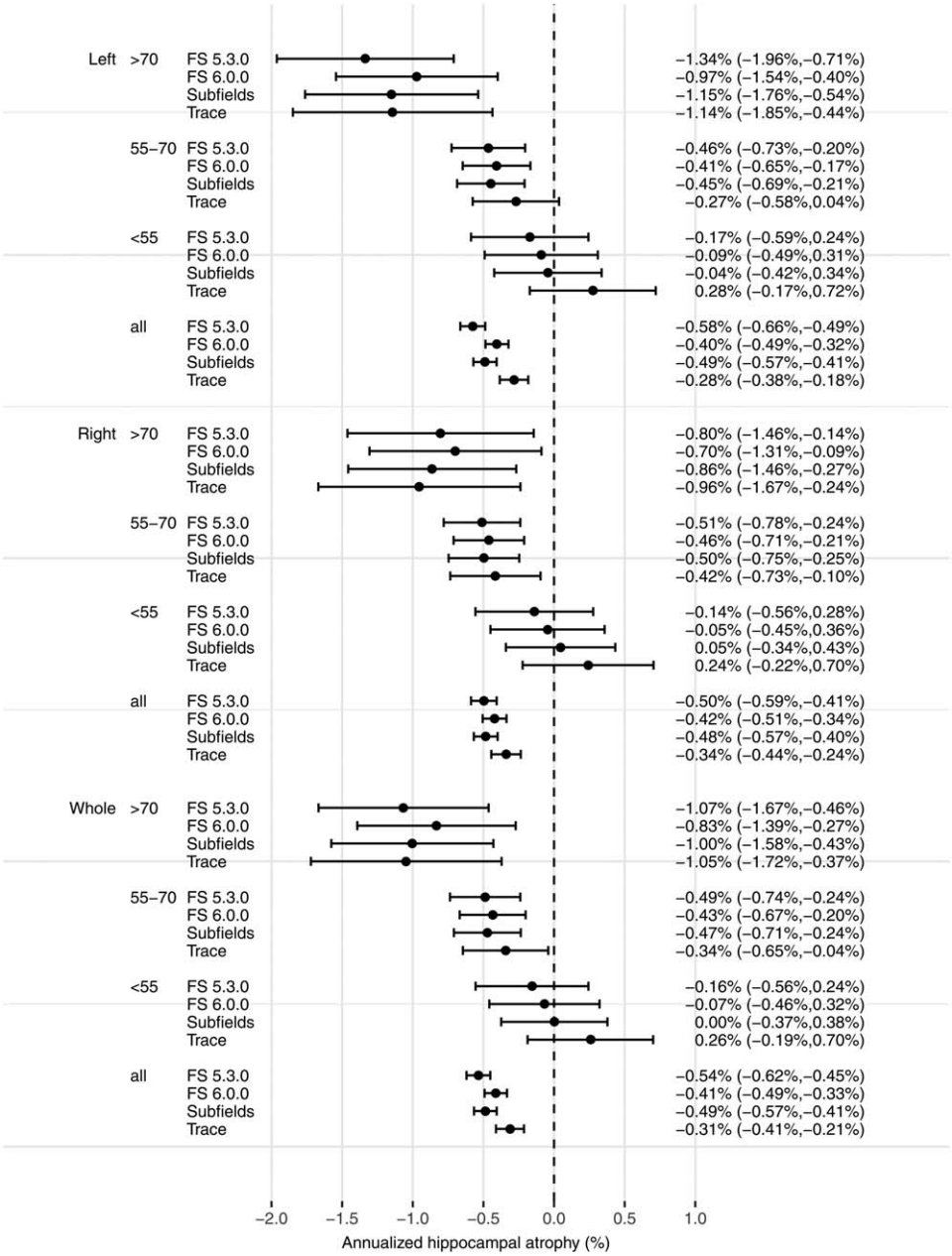

FIGURE 6 Quantitative forest plot of calculated annualized atrophy. Atrophy is reported as percentage of tissue lost each year, based on a generalized log-linear model of raw hippocampal volume over age in a cross-sectional sample. Atrophy is shallow or nonexistent in participants under 55 and greater in older strata, as expected, by all methods

of subcortical structures by multiple atlases and cortical metrics that would incur additional costs to replicate manually. Ideally, tools like FreeSurfer will continue to allow for cost-effective segmentations while periodic re-assessments of new versions allow for consistent interpretation of the published research findings based on them.

Concerns have been raised about FreeSurfer 5.3.0's overestimation of volume being relatively smaller in older participants (Wenger et al., 2014), which could skew age-related interpretations in older samples. Considering the wide use of FreeSurfer in cognitive aging studies, this represents a significant concern. After stratifying our sample by age, FreeSurfer 5.3.0 overestimated by 28% in those under 55, 23% in those between 55 and 70, and 21% in those over 70, qualitatively replicating Wenger et al.'s reduction of overestimation in the elderly (with smaller hippocampi). The same analysis fails to detect that bias in FreeSurfer 6.0.0, with overestimations of 23%, 19%, and 20%. Using a linear model to assess the overestimate over age revealed a coefficient of -23 mm^3 per year with version 5.3.0 and a lessened, but still significant, coefficient of -12 mm^3 per year with 6.0.0 (Figure 3). Consequently, calculations of annualized atrophy based on FreeSurfer 6.0.0 data match results from manual tracings more closely than older versions. (Figures 5 and 6, Table 5). Our sample's youngest participant was 38 when scanned, however, preventing us from precisely replicating Wenger et al.s 20–30 year-old sample. This improving, but still present age bias presents a statistical challenge not only when comparing

studies based on human tracing to those using FreeSurfer volumes, but also when comparing results using different FreeSurfer versions.

The right hippocampus is likely to be larger than the left, although some studies have reported the opposite (Pedraza et al., 2004). All methods used in this study support the consensus in reporting rightward (positive) asymmetry (Table 4), which is slightly less pronounced in left-handed participants. Previous studies of automated tools occasionally mention handedness, but none have had large enough samples to assess potential effects (Cherbuin et al., 2009; Wenger et al., 2014). This study assessed asymmetry in 40 left-handed and 620 right-handed participants, giving us more power to detect a handedness effect than any we've found to date. The differences in asymmetry index between right- and left-handed participants were still insignificant for all versions (e.g., manual tracing $p = .438$, and FreeSurfer 6.0.0 $p = .760$), suggesting no handedness effect, and that excluding left-handed participants from studies of hippocampal symmetry is unnecessary.

Relating hippocampal volumes to other variables often requires statistically adjusting volumes by TIV. This report includes both raw and TIV-adjusted volumes where appropriate for valid comparison with other literature. Manually-traced hippocampal volumes have been adjusted relative to manually-traced intracranial volumes. FreeSurfer segmentations have been adjusted relative to Freesurfer-generated TIV (eTIV in FreeSurfer files). This comparison likely reflects real-world scenarios where one complete pipeline would be selected over the other as a unit, and is therefore appropriate in the context of comparisons between methods. For studies investigating the relationship of hippocampal volume to another variable correlated with TIV, it may be more appropriate to adjust both hippocampal volume measures by the same TIV measure to reduce variance. Manually traced TIV correlated well with FreeSurfer TIV (version 6.0.0: $r = .84$; version 5.3.0: $r = .83$), indicating agreement between methods. Atrophy calculations were not significantly affected by adjustments for TIV (Figure 5a vs. b, Table 5), suggesting that atrophy may occur independent of original volume.

One idea to alleviate FreeSurfer's slight age bias is the creation and use of age-specific templates. This could result in more accurate individual segmentations, but is complicated by atlas selection biases and more difficulty in comparing results. Both FreeSurfer and human tracers in this study were blinded to all demographic information, including age and sex. Providing different templates to FreeSurfer based on age would remove that blind, and introduce an explicit age bias. Assuming bias could be dealt with statistically, determining whether a group of younger people, using young templates, differed from a group of older people, using elderly templates, would be impossible as the difference in templates introduces an alternate explanation for any differences found. One interesting possibility to address both problems may be if FreeSurfer could algorithmically create or select a template on-the-fly, based on a subset of existing templates, that more closely matches the subject under analysis. FreeSurfer would still be blinded to actual age, but may algorithmically select younger templates simply because the subject brain is more like them anatomically. This would remove explicit age-bias while still adjusting for apparent age and making the apparent-age-adjusted values part of the objective algorithm.

The objective of this study was to compare manual tracing to several methods and versions of FreeSurfer segmentation, using default settings and templates to represent the most common FreeSurfer use cases. To that end, our main comparison variables were from FreeSurfer's aseg-derived hippocampal volumes. Two alternate measures, sf+ and sf−, were also introduced to the analyses: sf+ simply to investigate FreeSurfer's improved hippocampal subfield routines (Iglesias et al., 2015); and sf− to bring FreeSurfer's segmentation boundaries closer to our manual tracing protocol. The difference between the smaller-than-expected whole_hippocampus subfield (sf+) and aseg-reported hippocampal volumes (Figure 2) resulted in sf+ having higher agreement with manual tracing than any other method tested. These results, even with similar biases to the aseg-based volumes (Figures 3 and 4), suggest the potential use of sf+ volumes rather than the aseg volumes in future studies focused on the hippocampus. This suggestion is further supported by improved Alzheimer disease discrimination using subfields' whole_hippocampus over aseg volumes (Iglesias et al., 2015).

This study has numerous limitations. It is limited by its narrow focus on the hippocampus and its selection of a single automated segmentation pipeline. FreeSurfer reports 45 subcortical volumes in addition to cortical volumes and areas from multiple atlases for potential comparison to manual segmentation methods. This study also focused on FreeSurfer to the exclusion of multiple other automated segmentation tools. The original 1.5T MRI acquisition parameters resulted in $0.9375 \times 0.9375 \times 1.6$ mm voxels, which were resampled to 1 mm isotropic resolution for FreeSurfer processing. The resampling has the potential to introduce biases to FreeSurfer's segmentation, relative to images originally obtained at 1mm isotropic resolution. The 1.5T field strength used here differs from the 3T strength being adopted in many present and future studies. Similar analyses with 3T MRI would provide valuable additions to this study as well as extending (Han et al., 2006) and (Iglesias et al., 2015). Moreover, our cohort represents only normal controls aged 38–84. Investigating these biases in younger people (Wenger et al., 2014) and people with disorders previously associated with hippocampal atrophy are important future directions.

A strength of this study is the collection of 664 MRI scans, complete with manual tracings and FreeSurfer segmentations. We are unaware of any comparably sized collection of manual tracings, particularly from a community sample covering males and females ranging so widely in age. This size allows sufficient statistical power to calculate atrophy in three independent age-based subgroups, without losing statistical significance. But it still does not contain enough left-handed participants to detect significant right > left asymmetry in left-handers alone. Power analyses report that 83 participants are necessary to have a 90% chance of detecting the volume biases we reported at a $p = .05$ level for FreeSurfer 6.0.0 or 43 participants for sf+. Smaller studies may be subject to these biases, but be underpowered to detect and report them, even with both automated and manually traced volumes.

Many older population-based studies, like GMBI or its related ARIC study, would benefit from reprocessing existing images with newer versions of FreeSurfer. Each study could provide additional

insights with a minimal expenditure of new funds. Having results from multiple field strengths would provide replication potential, richer information about the similarities and differences between field strengths, and a broader picture to feed meta-analyses. This report provides relevant information to investigators considering that approach and those who might seek to interpret their efforts in a larger context.

In summary, we found that FreeSurfer continues to report larger hippocampal volumes than manual tracing, in agreement with prior reports. The larger volumes correlate well with manual tracings, however, and show acceptable agreement in overlap analyses. Moreover, FreeSurfer and manual tracing report similar results in assessment of hippocampal atrophy over the lifespan, in line with prior studies (Fraser et al., 2015). Our results indicate that FreeSurfer is an adequate, although not directly comparable, replacement for manual tracing, provided manual assessment of input images and output segmentations for exclusionary criteria relevant to each investigation.

ACKNOWLEDGMENTS

The authors thank the GMBI study staff and participants for their contributions, and the Mississippi Center for Supercomputing Research for computing resources. We also thank Ben Downer for manual tracing assistance.

CONFLICTS OF INTEREST

All authors declare no conflict of interest. The sponsors played no role in the design, methods, subject recruitment, data collection, data analysis, or preparation of this manuscript.

ORCID

Mike F Schmidt http://orcid.org/0000-0003-4721-1457

REFERENCES

Alemán-Gómez, Y., Melie-Garcia, L., & Valdes-Hernandez, P. (2006). IBASPM: Toolbox for automatic parcellation of brain structures. In: Organization for Human Brain Mapping Vol. 27, p 2006. https://scholar.google.com/scholar?cluster=3093525772053243970.

Allen, J. S., Bruss, J., Brown, C. K., & Damasio, H. (2005). Normal neuroanatomical variation due to age: The major lobes and a parcellation of the temporal region. *Neurobiol Aging*, *26*(9), 1245–1260. http://linkinghub.elsevier.com/retrieve/pii/S0197458005001697.

Amazon Inc. (2017). EC2 Instance Pricing - Amazon Web Services (AWS). https://aws.amazon.com/ec2/pricing/on-demand/.

Arndt, S., Cohen, G., Alliger, R. J., Swayze, V. W., & Andreasen, N. C. (1991). Problems with ratio and proportion measures of imaged cerebral structures. *Psychiatry Res Neuroimaging*, *40*(1), 79–89. https://doi.org/10.1016/0925-4927(91)90031-K.

Barnes, J., Bartlett, J. W., van de Pol, L. A., Loy, C. T., Scahill, R. I., Frost, C., … Fox, N. C. (2009). A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. *Neurobiol Aging*, *30*(11), 1711–1723. http://linkinghub.elsevier.com/retrieve/pii/S0197458008000262.

Boccardi, M., Bocchetta, M., Apostolova, L. G., Barnes, J., Bartzokis, G., Corbetta, G., … Frisoni, G. B. (2015a). Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. *Alzheimer's Dement*, *11*(2), 126–113. https://doi.org/10.1016/j.jalz.2014.02.009.

Boccardi, M., Bocchetta, M., Ganzola, R., Robitaille, N., Redolfi, A., Duchesne, S., … Watson, C. (2015b). Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. *Alzheimer's Dement*, *11*(2), 184–194. https://doi.org/10.1016/j.jalz.2013.03.001.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., … Frisoni, G. B. (2011). Survey of protocols for the manual segmentation of the hippocampus: Preparatory steps towards a joint EADC-ADNI harmonized protocol. *Journal of Alzheimer's Disease*, *26 Suppl 3*, 61–75. https://doi.org/10.3233/JAD-2011-0004.

Campbell, S., Marriott, M., Nahmias, C., & MacQueen, G. M. (2004). Lower hippocampal volume in patients suffering from depression: A meta-analysis. *American Journal of Psychiatry*, *161*(4), 598–607. http://journals.psychiatryonline.org/article.aspx?articleid=176723.

Cherbuin, N., Anstey, K. J., R?Glade-Meslin, C., & Sachdev, P. S. (2009). In vivo hippocampal measurement and memory: A comparison of manual tracing and automated segmentation in a large community-based sample. Ed. Mark W. Greenlee. *PLoS One*, *4*(4), e5265. https://doi.org/10.1371/journal.pone.0005265.

Cook, M. J., Fish, D. R., Shorvon, S. D., Straughan, K., & Stevens, J. M. (1992). Hippocampal volumetric and morphometric studies in frontal and temporal lobe epilepsy. *Brain*, *115*(4), 1001–1015. http://brain.oxfordjournals.org/cgi/doi/10.1093/brain/115.4.1001.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Journal of Biomedical Informatics*, *29*(3), 162–173. http://linkinghub.elsevier.com/retrieve/pii/S0010480996900142.

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMRi N Biomedicine*, *10*(4–5), 171–178. http://afni.nimh.nih.gov/pub/dist/doc/papers/afni_paper2.pdf.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302. http://www.jstor.org/stable/1932409.

Du, A.-T., Schuff, N., Chao, L. L., Kornak, J., Jagust, W. J., Kramer, J. H., … Weiner, M. W. (2006). Age effects on atrophy rates of entorhinal cortex and hippocampus. *Neurobiol Aging*, *27*(5), 733–740. http://www.ncbi.nlm.nih.gov/pubmed/15961190.

Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2015). GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations. http://www.gnu.org/software/octave/doc/interpreter.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021.

Fischl, B., Salat, D. H., Busa, E., Albert, M. S., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. http://www.ncbi.nlm.nih.gov/pubmed/11832223.

Fraser, M. A., Shaw, M. E., & Cherbuin, N. (2015). A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. *Neuroimage*, *112*, 364–374. http://linkinghub.elsevier.com/retrieve/pii/S1053811915002177.

Frisoni, G. B., & Jack, C. R. (2011). Harmonization of magnetic resonance-based manual hippocampal segmentation: A mandatory step for wide clinical use. *Alzheimer's Dement*, *7*(2), 171–174. http://www.ncbi.nlm.nih.gov/pubmed/21414554.

Frisoni, G. B., Jack, C. R., Bocchetta, M., Bauer, C., Frederiksen, K. S., Liu, Y., … Boccardi, M. (2015). The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. *Alzheimer's Dement*, *11*(2), 111–125. http://linkinghub.elsevier.com/retrieve/pii/S1552526014024686.

Geuze, E., Vermetten, E., & Bremner, J. D. (2005). MR-based in vivo hippocampal volumetrics: 2. Findings in neuropsychiatric disorders. *Molecular Psychiatry*, *10*(2), 160–184. http://www.nature.com/doifinder/10.1038/sj.mp.4001579.

Gosche, K. M., Mortimer, J. A., Smith, C. D., Markesbery, W. R., & Snowdon, D. A. (2002). Hippocampal volume as an index of Alzheimer neuropathology: Findings from the Nun Study. *Neurology*, *58*(10), 1476–1482. http://www.ncbi.nlm.nih.gov/pubmed/12034782.

Gold, S., Christian, B., Arndt, S., Zeien, G., Cizadlo, T., Johnson, D. L., … Andreasen, N. C. (1998). Functional MRI statistical software packages: A comparative analysis. *Human Brain Mapping*, *6*, 73–84. https://doi.org/10.1002/(SICI)1097-0193(1998)6:2<73::AID-HBM1>3.0.CO;2-H.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., … Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *Neuroimage*, *32*(1), 180–194. https://www.sciencedirect.com/science/article/pii/S1053811906001601.

Han, X., & Fischl, B. (2007). Atlas renormalization for improved brain MR image segmentation across scanner platforms. *IEEE Transactions on Medical Imaging*, *26*(4), 479–486. http://www.ncbi.nlm.nih.gov/pubmed/17427735.

Hasan, K. M. (2009). A questionable gold standard for hippocampus volume and asymmetry. *Neuroradiology*, *51*(3), 201. 2–4. http://link.springer.com/10.1007/s00234-008-0492-5.

Hasan, K. M., & Pedraza, O. (2009). Improving the reliability of manual and automated methods for hippocampal and amygdala volume measurements. *Neuroimage*, *48*(3), 497–498. http://linkinghub.elsevier.com/retrieve/pii/S1053811909004984.

Hasboun, D., Chantôme, M., Zouaoui, A., Sahel, M., Deladoeuille, M., Sourour, N., … Dormont, D. (1996). MR determination of hippocampal volume: Comparison of three methods. *American Journal of Neuroradiology*, *17*, 1091–1098. http://www.ajnr.org/content/17/6/1091.

Hosseini, M. P., Nazem-Zadeh, M. R., Pompili, D., Jafari-Khouzani, K., Elisevich, K., & Soltanian-Zadeh, H. (2016). Comparative performance evaluation of automated segmentation methods of hippocampus from magnetic resonance images of temporal lobe epilepsy patients. *Medical Physics*, *43*(1), 538–553. http://doi.wiley.com/10.1118/1.4938411.

Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., … Van Leemput, K. (2015). A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. *Neuroimage*, *115*, 117–137. https://doi.org/10.1016/j.neuroimage.2015.04.042.

Jack, C. R. (1994). MRI-based hippocampal volume measurements in epilepsy. *Epilepsia*, *35*(s6), S21–S29. http://doi.wiley.com/10.1111/j.1528-1157.1994.tb05986.x.

Jack, C. R., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., … Kokmen, E. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology*, *51*(4), 993–999. https://www.neurology.org/content/51/4/993.full.

Jack, C. R., Sharbrough, F. W., Twomey, C. K., Cascino, G. D., Hirschorn, K. A., Marsh, W. R., … Scheithauer, B. (1990). Temporal lobe seizures: Lateralization with MR volume measurements of the hippocampal formation. *Radiology*, *175*(2), 423–429. https://doi.org/10.1148/radiology.175.2.2183282.

Jack, C. R., Theodore, W. H., Cook, M. J., & McCarthy, G. (1995). MRI-based hippocampal volumetrics: Data acquisition, normal ranges, and optimal protocol. *Magnetic Resonance Imaging*, *13*(8), 1057–1064. http://www.sciencedirect.com/science/article/pii/0730725X9502013J.

Jack, C. R., Twomey, C. K., Zinsmeister, A. R., Sharbrough, F. W., Petersen, R. C., & Cascino, G. D. (1989). Anterior temporal lobes and hippocampal formations: Normative volumetric measurements from MR images in young adults. *Radiology*, *172*(2), 549–554. http://pubs.rsna.org/doi/abs/10.1148/radiology.172.2.2748838.

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., … Blacker, D. (2009). MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage*, *46*(1), 177–192. https://doi.org/10.1016/j.neuroimage.2009.02.010.

Kesslak, J. P., Nalcioglu, O., & Cotman, C. W. (1991). Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer's disease. *Neurology*, *41*(1), 51–54. http://www.neurology.org/content/41/1/51.short.

Knopman, D. S., Mosley, T. H., Bailey, K. R., Jack, C. R., Schwartz, G. L., & Turner, S. T. (2008). Associations of microalbuminuria with brain atrophy and white matter hyperintensities in hypertensive sibships. *Journal of Neurological Sciences*, *271*(1–2), 53–60. http://linkinghub.elsevier.com/retrieve/pii/S0022510X08001457.

Konrad, C., Ukas, T., Nebel, C., Arolt, V., Toga, A. W., & Narr, K. L. (2009). Defining the human hippocampus in cerebral magnetic resonance images—An overview of current segmentation protocols. *Neuroimage*, *47*(4), 1185–1195. https://doi.org/10.1016/j.neuroimage.2009.05.019.

Li, Y.-J., Ga, S.-N., Huo, Y., Li, S.-Y., & Gao, X.-G. (2007). Characteristics of hippocampal volumes in healthy Chinese from MRI. *Neurological Research*, *29*(8), 803–806. http://www.tandfonline.com/doi/full/10.1179/016164107X223557.

Logue, M. W., van Rooij, S. J., Dennis, E. L., Davis, S. L., Hayes, J. P., Stevens, J. S., … Morey, R. A. (2017). Smaller hippocampal volume in posttraumatic stress disorder: A multi-site ENIGMA-PGC study. *Biological Psychiatry*, *83*, 244–253. http://linkinghub.elsevier.com/retrieve/pii/S000632231731990X.

Makris, N., Kennedy, D. N., Meyer, J., Worth, A., Caviness, V. S., Jr Seidman, L. J., … Boriel, D. Sanders H. (2004, May). CMA methodology overview. Retrieved from http://www.cma.mgh.harvard.edu/manuals/segmentation/.

McAuliffe, M. J., Lalonde, F. M., McGarry, D., Gandler, W., Csaky, K., & Trus, B. L. (2001). Medical image processing, analysis and visualization in clinical research. In: Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001. IEEE Comput. Soc. pp 381–386. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=941749.

Melton, L. J. (1996). History of the rochester epidemiology project. *Mayo Clinic Proceedings*, *71*(3), 266–274. http://linkinghub.elsevier.com/retrieve/pii/S0025619611639669.

Moghaddam, M. J., & Soltanian-Zadeh, H. (2009). Automatic segmentation of brain structures using geometric moment invariants and artificial neural networks. *Information Processing in Medical Imaging*, *21*, 326–337. http://link.springer.com/10.1007/978-3-642-02498-6_27.

Morey, R. A., Petty, C. M., Xu, Y., Pannu Hayes, J., Wagner, H. R., Lewis, D. V., … McCarthy, G. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *Neuroimage*, *45*(3), 855–866. https://doi.org/10.1016/j.neuroimage.2008.12.033.

Morey, R. A., Selgrade, E. S., Wagner, H. R., Huettel, S. A., Wang, L., & McCarthy, G. (2010). Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Human Brain Mapping*, *31*, 1751–1762. http://doi.wiley.com/10.1002/hbm.20973.

NIH (2016). NOT-OD-16-047. Ruth L. Kirschstein National Research Service Award (NRSA) Stipends, Tuition/Fees and Other Budgetary Levels Effective for Fiscal Year 2016. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-047.html.

Pedraza, O., Bowers, D., & Gilmore, R. (2004). Asymmetry of the hippocampus and amygdala in MRI volumetric measurements of normal adults. *Journal of the International Neuropsychological Society*, *10*(05), 664–678. http://www.journals.cambridge.org/abstract_S1355617704105080.

Platero, C., & Tobar, M. C. (2016). A fast approach for hippocampal segmentation from T1-MRI for predicting progression in Alzheimer's disease from elderly controls. *Journal of Neuroscience Methods*, *270*, 61–75. https://doi.org/10.1016/j.jneumeth.2016.06.013.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raz, N., Gunning-Dixon, F., Head, D., Rodrigue, K. M., Williamson, A., & Acker, J. D. (2004a). Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: Replicability of regional differences in volume. *Neurobiology Aging*, *25*, 377–396. http://www.ncbi.nlm.nih.gov/pubmed/15123343.

Raz, N., & Rodrigue, K. M. (2006). Differential aging of the brain: Patterns, cognitive correlates and modifiers. *Neuroscience & Biobehavioral Reviews*, *30*(6), 730–748. http://www.ncbi.nlm.nih.gov/pubmed/16919333.

Raz, N., Rodrigue, K. M., Head, D., Kennedy, K. M., & Acker, J. D. (2004). Differential aging of the medial temporal lobe: A study of a five-year change. *Neurology*, *62*(3), 433–438. http://www.neurology.org/cgi/doi/10.1212/01.WNL.0000106466.09835.46.

Revelle, W. (2017). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, version 1.7.8. https://CRAN.R-project.org/package=psych.

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 Research*, *181*(3), 219–225. http://www.ncbi.nlm.nih.gov/pubmed/20153146.

Schmidt, M. F. (2017). freesurfer-management code on github. https://github.com/mfschmidt/freesurfer-management.

Schoemaker, D., Buss, C., Head, K., Sandman, C. A., Davis, E. P., Chakravarty, M. M., … Pruessner, J. C. (2016). Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *Neuroimage*, *129*, 1–14. https://doi.org/10.1016/j.neuroimage.2016.01.038.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Smith, J. A., Turner, S. T., Sun, Y. V., Fornage, M., Kelly, R. J., Mosley, T. H., … Kardia, S. L. (2009). Complexity in the genetic architecture of leukoaraiosis in hypertensive sibships from the GENOA Study. *BMC Medical Genomics*, *2*(1), 16. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2679055&tool=pmcentrez&rendertype=abstract.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., … Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208–S219. http://www.sciencedirect.com/science/article/pii/S1053811904003933.

Squire, L. R., Stark, C. E. L., & Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neuroscience*, *27*(1), 279–306. http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.27.070203.144130.

Tae, W. S., Kim, S. S., Lee, K. U., Nam, E.-C., & Kim, K. W. (2008). Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. *Neuroradiology*, *50*(7), 569–581. http://link.springer.com/10.1007/s00234-008-0383-9.

Tae, W. S., Kim, S. S., Lee, K. U., Nam, E.-C., & Kim, K. W. (2009). Validation of hippocampal volumes measured with one manual and two automated methods using FreeSurfer and IBASPM in chronic major depressive disorder. *Neuroradiology*, *51*(3), 203–204. http://link.springer.com/10.1007/s00234-008-0492-5.

Tangaro, S., Amoroso, N., Boccardi, M., Bruno, S., Chincarini, A., Ferraro, G., … Bellotti, R. (2014). Automated voxel-by-voxel tissue classification for hippocampal segmentation: Methods and validation. *Physica Medica*, *30*(8), 878–887. https://doi.org/10.1016/j.ejmp.2014.06.044.

Turner, J. (2014). The rise of large-scale imaging studies in pschiatry. *Gigascience*, *3*, 29. https://doi.org/10.1186/2047-217X-3-29.

Turner, S. T., Fornage, M., Jack, C. R., Mosley, T. H., Kardia, S. L. R., Boerwinkle, E., & de Andrade, M. (2005). Genomic susceptibility loci for brain atrophy in hypertensive sibships from the GENOA study. *Hypertension*, *45*(4), 793–798. http://www.ncbi.nlm.nih.gov/pubmed/15699467.

Videbech, P., & Ravnkilde, B. (2004). Hippocampal volume and depression: A meta-analysis of MRI studies. *The American Journal of Psychiatry*, *161*(11), 1957–1966. http://journals.psychiatryonline.org/article.aspx?articleid=177136.

Walhovd, K. B., Fjell, A. M., Reinvang, I., Lundervold, A., Dale, A. M., Eilertsen, D. E., … Fischl, B. (2005). Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology Aging*, *26*(9), 1261–1270. https://doi.org/10.1016/j.neurobiolaging.2005.05.020.

Walhovd, K. B., Westlye, L. T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., … Fjell, A. M. (2011). Consistent neuroanatomical age-related volume differences across multiple samples. *Neurobiology Aging*, *32*(5), 916–932. https://doi.org/10.1016/j.neurobiolaging.2009.05.013.

Watson, C., Andermann, F., Gloor, P., Jones-Gotman, M., Peters, T., Evans, A., … Leroux, G. (1992). Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. *Neurology*, *42*(9), 1743–1750. https://doi.org/10.1212/WNL.42.9.1743.

Wenger, E., Mårtensson, J., Noack, H., Bodammer, N. C., Kühn, S., Schaefer, S., … Lövdén, M. (2014). Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains. *Human Brain Mapping*, *35*(8), 4236–4248. http://doi.wiley.com/10.1002/hbm.22473.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Schmidt MF, Storrs JM, Freeman KB, et al. A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan. *Hum Brain Mapp*. 2018;39:2500–2513. https://doi.org/10.1002/hbm.24017