Brain Topography
https://doi.org/10.1007/s10548-018-0659-2

ORIGINAL PAPER

Longitudinal Assessment of Hippocampal Atrophy in Midlife and Early Old Age: Contrasting Manual Tracing and Semi-automated Segmentation (FreeSurfer)

Mark A. Fraser[1] · **Marnie E. Shaw**[2] · **Kaarin J. Anstey**[1] · **Nicolas Cherbuin**[1]

Received: 14 March 2018 / Accepted: 29 June 2018

Abstract

It is important to have accurate estimates of normal age-related brain structure changes and to understand how the choice of measurement technique may bias those estimates. We compared longitudinal change in hippocampal volume, laterality and atrophy measured by manual tracing and FreeSurfer (version 5.3) in middle age (n = 244, 47.2[1.4] years) and older age (n = 199, 67.0[1.4] years) individuals over 8 years. The proportion of overlap (Dice coefficient) between the segmented hippocampi was calculated and we hypothesised that the proportion of overlap would be higher for older individuals as a consequence of higher atrophy. Hippocampal volumes produced by FreeSurfer were larger than manually traced volumes. Both methods produced a left less than right volume laterality difference. Over time this laterality difference increased for manual tracing and decreased for FreeSurfer leading to laterality differences in left and right estimated atrophy rates. The overlap proportion between methods was not significantly different for older individuals, but was greater for the right hippocampus. Estimated middle age annualised atrophy rates were −0.39(1.0) left, 0.07(1.01) right, −0.17(0.88) total for manual tracing and −0.15(0.69) left, −0.20(0.63) right, −0.18(0.57) total for FreeSurfer. Older age atrophy rates were −0.43(1.32) left, −0.15(1.41) right, −0.30 (1.23) total for manual tracing and −0.34(0.79) left, −0.68(0.78) right, −0.51(0.65) total for FreeSurfer. FreeSurfer reliably segments the hippocampus producing atrophy rates that are comparable to manual tracing with some biases that need to be considered in study design. FreeSurfer is suited for use in large longitudinal studies where it is not cost effective to use manual tracing.

Keywords Hippocampus · Longitudinal · FreeSurfer · Manual tracing · Normal ageing · Magnetic resonance imaging

Introduction

Hippocampal volumes change across the adult lifespan (Raz et al. 2010) as a result of the mental demands of acquiring knowledge (Draganski et al. 2006; Woollett and Maguire 2011), physical activity (Erickson et al. 2011), ageing (Fjell et al. 2013; Jinno 2015), chronic disease and neurodegenerative conditions. Historically, hippocampal volumes have been measured manually by tracing the outline of the hippocampus slice by slice on MRI scans, and more recently by automated and semi-automated methods. The hippocampus is a challenging structure to delineate on MRI due to its irregular shape, indistinct border with the amygdala and high degree of individual variability. Manual tracing is considered the gold standard for accuracy (Barnes et al. 2009b; Boccardi et al. 2011) in measuring hippocampal volumes in-vivo. However, manual tracing requires a high level of skill and it can take several hours to process each scan. Reliability

Handling editor: Andrew Zalesky.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10548-018-0659-2) contains supplementary material, which is available to authorized users.

✉ Mark A. Fraser
mark.fraser@anu.edu.au

[1] Centre for Research on Ageing, Health and Wellbeing, Australian National University, Florey, Building 54, Mills Road, Canberra, ACT 2601, Australia

[2] College of Engineering & Computer Science, Australian National University, Brian Anderson Building 115, 115 North Road, Canberra, ACT 2601, Australia

Published online: 04 July 2018

between raters and across time needs to be managed and raters may be affected by biases such as visual laterality bias (Boccardi et al. 2011; Maltbie et al. 2012; Rogers et al. 2012). As a result, few large studies use manual tracing due to the high cost in time and resources. In contrast, automated segmentation methods have the advantages of being able to process large MRI datasets with little human intervention, allow for the reprocessing of scans when more accurate versions of the software become available, and for some packages, segment all structures of the brain simultaneously (Cash et al. 2015; Dill et al. 2015). Consequently, the choice of measurement technique depends upon one's requirements for accuracy, repeatability and the size of the study.

There have been many studies that compare manual tracing to a variety of automated segmentation approaches. However, most used a cross-sectional design and focused on subjects under 40 or over 60 years of age with relatively few including middle age samples (Cherbuin et al. 2009; de Flores et al. 2014). Cross-sectional comparisons between manual and automated segmentation suggest that both approaches can reliably detect differences in hippocampal volume resulting from a range of pathological factors including volume differences between controls and individuals with acute depression (Bergouignan et al. 2008), chronic depressive disorder (Tae et al. 2008), schizophrenia or schizoaffective disorder (Arnold et al. 2015), bipolar affective disorder (Doring et al. 2011), as well as between healthy persons and those with cognitive impairment or Alzheimer's Dementia (AD: Cash et al. 2015, Hsu et al. 2002; Leung et al. 2010; Shen et al. 2010). Yet, in order to better identify and understand the impact of chronic disease and neurodegenerative conditions, it is important to have accurate estimates of normal age related brain structure changes. Longitudinal designs, where each individual acts as their own control, are required to accurately measure hippocampal change over time (Pfefferbaum and Sullivan 2015; Raz et al. 2010; Scahill et al. 2003) and in the few longitudinal comparison studies published to date, both manual and automated approaches show that atrophy increases as cognitive impairment increases (Barnes et al. 2008, 2004; Leung et al. 2010; Mulder et al. 2014).

Further investigation is needed to better understand whether different methodologies produce different results and to what extent estimates are affected by the age of the sample surveyed or other factors. Firstly, it is not known whether the differences in the segmented hippocampi, the volumetric estimates or the atrophy rates between manual tracing and automated segmentation are consistent for middle aged and older individuals. It is important to be able to accurately measure small hippocampal changes in midlife that may represent the prodromal stages of diseases such as AD (Braak et al. 1996; Fjell et al. 2013; Frisoni et al. 2010). Systematic differences between young individuals and old individuals have been found (Wenger et al. 2014), but possible differences between midlife individuals, with little atrophy, and older individuals with greater levels of atrophy are yet to be tested. We hypothesised that the difference between manually traced and FreeSurfer segmented hippocampi, in terms of the proportion of overlap of the segmented voxels, would be less pronounced in the older sample as a consequence of higher atrophy in older individuals increasing the proportion of high contrast hippocampal boundaries adjacent to areas of CSF being more easily delineated by both manual and automated methods. Secondly, it is unclear whether automated segmentation methods produce similar estimates of laterality to manual tracing as the few studies to compare laterality between manual tracing and automated segmentation, using FreeSurfer, have produced equivocal results (Sánchez-Benavides et al. 2010; Tae et al. 2008; Wenger et al. 2014).

In this study, we compared FreeSurfer segmentation to manual tracing in large middle age (44–48 years) and older age (64–68 years) population samples over a period of eight years to investigate the impact of segmentation method and the relative age of the samples on estimated volumes, overlap of the segmented hippocampi, laterality, and atrophy.

Materials and Method

Participants

Participants were selected from the Personality and Total Health through life (PATH) project, a large longitudinal study investigating mood disorders, substance use, cognition, health indicators and individual characteristics across the adult lifespan (Anstey et al. 2012). At baseline, the PATH project surveyed 2404 young individuals aged 20–24 years, 2530 middle age (MA) individuals aged 40–44 years and 2551 older age (OA) individuals aged 60–64 years, who were residents of the adjacent cities of Canberra and Queanbeyan, Australia. The participants were randomly recruited through the electoral roll and as enrolment to vote is compulsory in Australia, this approach provided a cohort that is representative of the population. Since the projects inception in 1999, there have been four waves of data collected for each age cohort, with each wave separated by 4 years. A subset of the two older groups, the MRI Sub-study, was randomly recruited to undergo additional tests that included MRI scans and blood tests. At wave 1, a subsample of 478 participants from the OA group underwent a baseline MRI scan and at the second wave of data collection 431 participants of the MA group received a baseline MRI scan. Participants in the MRI sub study were then invited to have further scans at each subsequent wave of data collection. All participants provided written informed consent. The Human Research

Ethics Committee of The Australian National University approved the study protocol. Further details of the survey including the sampling procedure and the MRI sub-study are reported elsewhere (Anstey et al. 2012; Cherbuin et al. 2009).

The current study focuses on data collected in the second and fourth waves of the MRI sub-study for the MA and OA groups where manually traced MRI scans were available for both the groups. Participants with stroke, epilepsy, Parkinson's disease, mild cognitive impairment (MCI) or dementia at any wave were excluded (Supplementary Table S1). For clarity PATH Wave 2 is referred to as Time-1 and PATH Wave 4 is referred to as Time-2 hereon. All individuals who had a valid manually traced scan and a successfully segmented FreeSurfer scan at both time points were included in the study (Table 1). There were no significant differences between the overall PATH sample and the study samples in terms of hypertension, diabetes, *APOE-ε4* genotype and current smoking status. The MA sample was older by 0.26 CI (0.07, 0.44) years t(319) = 2.69, p = 0.007 at Time-2. The OA sample was 0.58 CI (0.21, 0.96) years more educated t (240) = 3.04, p = 0.003, had a lower BMI by 0.64 CI (1.22, 0.06) years t(248) = −2.16, p = 0.031, and had a higher proportion of males, $\chi 2$ (1) = 4.31, p = 0.038, than the PATH sample at Time-1. At Time-2, the OA sample was −0.48 CI (−0.68, −0.28) years younger t(248) = −4.76, p < 0.001, was 0.40 CI (0.01, 0.78) years more educated t (254) = 2.02, p = 0.045, and had a higher proportion of males, $\chi 2$ (1) = 3.85, p = 0.050.

MRI Scan Acquisition

T1-weighted three-dimensional structural MRI scans were acquired for all participants using 1.5T MRI scanners. Images were obtained at Time-1 with a Phillips Gyroscan ACS-NT scanner (Phillips Medical systems, Best, The Netherlands) in coronal orientation using a fast-field echo sequence with the repetition time (TR), echo time (TE) and flip angle of 8.93 ms/3.57 ms/8^0, slice thickness of 1.5 mm, and matrix size of 256 × 256 giving a voxel size of 1.5mm^3. At Time-2, scans were acquired on a Siemens Espree scanner (Siemens Medical solutions) in sagittal orientation using a MPRAGE sequence with TR, TE and flip angle of 1160 ms/4.24 ms/15^0, slice thickness of 1 mm, and matrix size 512 × 512 giving a voxel size of 0.25mm^3.

Table 1 Demographic characteristics of the study samples for each age group and time point

	MA		OA	
	Time-1	Time-2	Time-1	Time-2
N	244	244	199	199
Age (SD)	47.2 (1.4)	55.7 (1.4)	67.0 (1.4)	75.1 (1.3)
Males (%)	115 (47.1)	115 (47.1)	118 (59.3)	118 (59.3)
Education(SD)	14.8 (2.2)	14.9 (2.2)	14.5 (2.6)	14.5 (2.6)
BMI (SD)	27.1 (4.7)	27.8 (5.0)	26.0 (3.9)	26.4 (4.2)
Hypertension (%)	65 (26.6)	97 (39.8)	129 (64.8)	156 (78.4)
Diabetes (%)	3 (1.2)	11 (4.5)	20 (10.1)	32 (16.1)
APOE-ε4 (%)	61 (25.0)	61 (25.0)	49 (24.6)	49 (24.6)
Smokers (%)	35 (14.3)	24 (9.8)	11 (5.5)	9 (4.5)

MA middle age, *OA* older age, *Education* years of education, *BMI* Body mass index, *APOE-ε4* participants with at least one *APOE-ε4* allele, *Smokers* current smokers

Image Pre-processing

The MINC imaging toolbox (MINC; http://en.wikibooks.org/wiki/MINC) was used for image intensity normalisation and B1 inhomogeneity correction (Sled et al. 1998) on all images. The starting point for both segmentation methods were these pre-processed images.

Image Segmentation

Manual Tracing

Manual hippocampal volumes were measured by tracing the hippocampi on each slice of a T1-weighted scan in coronal orientation using Analyze Software (Brain Imaging Resource, Mayo Clinic, Rochester, MI, USA). The outlining of the hippocampus proceeded in an anterior to posterior direction with the head and body traced according to the protocol outlined by Watson et al. (1997), and the hippocampal tail according to the protocol of Maller et al. (2006). In addition, areas of cerebrospinal fluid (CSF) wholly enclosed in the hippocampus were excluded from the hippocampal region of interest (ROI;Maller et al. 2011). During tracing, the images were displayed in radiological orientation, with the right hippocampus on left side of screen, and the right hippocampus was traced first. The images from Time-2 were resized to 1mm^3 voxels prior to tracing. All scans used in this study were traced by a single highly experienced neurologist (CM) at the time of data collection. For each time point and age group, the scans of 10 individuals were re-traced to compute an intra-class correlation (ICC) measure and the ICCs were > 0.97 for all measures. While an inter-class correlation measure was not computed for the samples in this study, such a measure was computed (using the same rater) for an earlier sample of the PATH study that demonstrated very high inter-rater reliability > 0.95 (Cherbuin et al. 2009, 2014).

Automated Segmentation

The volumetric segmentation was performed on the MRI scans from all four waves of the PATH project, using the FreeSurfer version 5.3 image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). The FreeSurfer cross-sectional pipeline converted the original scans to $256 \times 256 \times 256$ $1mm^3$ isotropic images and transformed them to Talairach space prior to segmentation being performed. The segmentation process involved labelling each voxel based on probabilistic information using both tissue intensity and location based on a manually labelled atlas (Fischl et al. 2002). The segmented subcortical ROIs generated by FreeSurfer were then corrected for partial volume effects (Gronenschild et al. 2012). Quality control of Freesurfer processing was implemented by an in-house script that identified possible outlier scans based on total grey and white matter volumes. The segmentations of the potential outliers were then visually checked to identify those that had failed FreeSurfer processing resulting in 24 scans being excluded as outliers: MA [Time-1 = 11 (2.6%), Time-2 = 6 (2.2%)] and OA [Time-1 = 2 (0.6%), Time-2 = 5 (1.8%)].

Statistical Analysis

Statistical analysis was performed using Matlab R2014b or R version 3.3.3 running under RStudio version 1.0.136.

Study Sample

The demographic characteristics of the participants used for this study were compared to the characteristics of the overall PATH study using t-tests and χ^2 tests of difference to establish any significant differences between the study samples and the overall PATH population samples.

The manual tracing and FreeSurfer hippocampal volumes were first normalized to adjust for differences in total intracranial volume (ICV) using the formula:

$$Vol_{adj} = Vol_{raw} - \beta \times (ICV - \overline{ICV}) \quad (1)$$

where β is the slope of the regression of the ROI volume on the ICV

The volumes were normalized using the estimated total intracranial volume for each participant calculated from the FreeSurfer cross-sectional pipeline (Buckner et al. 2004). Because the scanner and MRI parameters changed between waves, we orthogonalised the volume data across the four PATH waves with respect to a scanner covariate to remove the scanner specific variance using a published and extensively scrutinised methodology (Fjell et al. 2013; Shaw et al. 2016a, 2017, b; Walsh et al. 2017). The distributions of the hippocampal volumes for each time point, age group and segmentation method were then examined and univariate outliers ($Z > 3.29$) were excluded. The record counts for each stage of the exclusion process are shown in Supplementary Table S1.

Longitudinal Analysis

Longitudinal analysis was then performed for individuals where both manually traced volumes and volumes segmented by the FreeSurfer pipeline were available, at both time points, after univariate outliers had been removed. The annualised rates of hippocampal atrophy were then calculated as follows:

$$A_{vol} = \frac{(Vol_{t2} - Vol_{t1})}{(Age_{t2} - Age_{t1})} \quad (2)$$

$$A_{\%} = \frac{A_{vol}}{Vol_{t1}} \times 100 \quad (3)$$

where A_{vol} is the annualised atrophy volume in mm^3 and $A_{\%}$ is the percentage annualised atrophy.

The distributions of atrophy rates for each age group and segmentation method were examined and univariate outliers were excluded. The manual and FreeSurfer atrophy rates were then compared using Bland Altman plots with the mean bias and limits of agreement (LoA) calculated in each case. The regression method for calculating the LoAs was used where there was a non-constant bias between the methods (Carstensen 2010). The effect of segmentation method, age group, gender and laterality on atrophy was examined using generalised linear models (GLM).

Sensitivity Analysis

While manual tracing is considered the gold standard for measuring hippocampal volumes, there are a range of manual tracing protocols, some that include the hippocampal tail and some that do not. In the PATH study, each hippocampi was traced as two separate ROIs with the first containing the hippocampal head and body traced according to the protocol of Watson et al. (1997) that excluded parts of the hippocampal tail anterior to the crus of the fornix. The second ROI, containing the hippocampal tail, was traced according to the protocol of Maller et al. (2006). Having the body and tail traced as separate ROIs enabled us to perform sensitivity analysis to investigate the effect of including or excluding the hippocampal tail on atrophy rates. Sensitivity analysis was also performed to investigate the impact on atrophy estimates of excluding individuals who had not been diagnosed

with MCI or dementia, but had scored less than 27 on the MMSE test. Low MMSE scores may indicate the prodromal stage of neuropathology.

Laterality Analysis

A laterality index was calculated to determine whether manual tracing and FreeSurfer performed differently in estimating left and right volumes using the following formula:

$$Laterality = \frac{(Vol_{left} - Vol_{right})}{(Vol_{left} + Vol_{right})} \times 100 \quad (4)$$

Similarity Analysis

The similarity between FreeSurfer and manual tracing segmentation was investigated using voxel by voxel comparisons of the baseline segmented scans. Similarity was measured using the Dice Coefficient (Dice 1945) that measures the proportion of overlap between the manually traced ROIs and the FreeSurfer segmented ROIs. The Dice coefficient is calculated with the following formula.

$$DSC(A, B) = \frac{2(A \cap B)}{(A + B)} \quad (5)$$

where A is the set of voxels included in the manual tracing ROI, and B is the set of voxels included in the FreeSurfer ROI.

Higher Dice coefficients indicate closer correspondence between the segmentation methods where a Dice coefficient of 1.0 would indicate complete overlap and a Dice coefficient of 0.0 would indicate that none of the voxels of either structure overlap.

To enable overlap comparisons to be performed, the FreeSurfer sub-cortical segmentation label files were first registered to the scans used for manual tracing using the label2vol utility (https://surfer.nmr.mgh.harvard.edu/fswiki/FsAnat-to-NativeAnat). This had the effect of transforming the FreeSurfer segmentation labels to be in the same location space and to have the same dimensions as the manually traced labels. The Dice coefficients were then calculated and a merged set of segmentations was created with labels to indicate overlapping (true positive), manual only (false negative), or FreeSurfer only (false positive) voxels. Any scan pairs with a Dice coefficient of <0.6 were visually checked to identify possible errors, such as segmentation errors or failed registration, requiring exclusion.

The resulting merged segmentations were examined to identify patterns of similarities and differences between manual tracing and FreeSurfer. First the merged segmentations of each hippocampi on all scans were visually reviewed by examining a single slice in each orientation (axial, sagittal and coronal) to assess the patterns of differences and similarities between the segmentation methods. Then the segmentations with the highest dice coefficients in each sample were inspected in detail using the Freeview package (https://surfer.nmr.mgh.harvard.edu/fswiki/ToolsTutorial#Freeview) to identify additional differences between manual tracing and FreeSurfer. Finally, GLMs were used to analyse the sources of variance in the resulting Dice coefficients.

Results

The MA sample was composed of 115 men and 129 women while the OA sample was composed of 118 men and 81 women. The demographic characteristics for the study are shown in Table 1.

Longitudinal Analysis

The distributions of the hippocampal volumes for each time point, age group and segmentation method were examined with six MA and five OA outliers excluded. The resulting mean volumes are shown in Table 2.

Annualised Atrophy rates were calculated for manual tracing and FreeSurfer. The distributions of the atrophy

Table 2 Mean hippocampal volumes for manual tracing and FreeSurfer

Age group	Method	Time	Hippocampal volume		
			n	Left (SD)	Right (SD)
MA	Manual	Time-1	238	3398.57 (326.19)	3463.12 (304.35)
		Time-2	238	3279.13 (311.17)	3478.69 (326.86)
	FreeSurfer	Time-1	238	4248.75 (343.76)	4323.13 (349.99)
		Time-2	238	4191.17 (369.78)	4243.91 (347.42)
OA	Manual	Time-1	194	3192.99 (312.39)	3321.53 (317.48)
		Time-2	194	3072.99 (356.12)	3276.36 (404.73)
	FreeSurfer	Time-1	194	3864.07 (315.51)	4021.17 (283.37)
		Time-2	194	3741.20 (363.52)	3789.75 (362.64)

MA middle age, *OA* older age

rates for each age group and segmentation method were examined with eight MA and four OA outliers excluded.

As shown in Table 3, compared to manual tracing, FreeSurfer estimated atrophy rates were lower for the left hippocampus, higher for the right hippocampus and similar for the total hippocampus in MA individuals. FreeSurfer estimated atrophy rates were similar for the left hippocampus, higher for the right hippocampus and higher for the total hippocampus in OA individuals. The variance of the FreeSurfer estimates were consistently lower than those for manual tracing. There were some laterality differences as FreeSurfer mean atrophy was greater for the right hippocampus than the left hippocampus in the OA individuals, whereas for manual tracing, atrophy was greater for the left hippocampus than the right in MA individuals. The BA plots of annualised atrophy rates show that the relationship between the atrophy rates for manual tracing and FreeSurfer were not consistent (Bland and Altman 1999; Carstensen 2010) with the difference between the atrophy rates for the two methods increasing as the absolute magnitude of the average atrophy rates increased (Fig. 1).

GLM models to investigate the impact of method, age, age group, gender and laterality on percentage atrophy rate indicated that FreeSurfer atrophy rates were significantly lower than manual tracing for the left hippocampus and significantly higher than manual tracing for the right hippocampus. Manual tracing atrophy rates were significantly lower for the right hippocampus compared to the left hippocampus. Compared to MA individuals, FreeSurfer atrophy estimates were significantly higher for the OA group while for manual tracing there was a non-significant trend for greater atrophy estimates in OA individuals (Table 4).

Sensitivity Analysis

Repeating the longitudinal analysis with manual tracing volumes based on the hippocampal body excluding the tail had limited effect on the associations between manual tracing and FreeSurfer atrophy rates. The non-significant effects for age and older individuals became significant and the significant interaction between FreeSurfer and OA individuals became a non-significant trend in the GLM model that excluded the hippocampal tail. Results from the sensitivity analysis excluding the hippocampal tail are shown in Supplementary Tables S2, S3, and Supplementary Fig. S1.

Sensitivity analysis to investigate the impact of excluding nine OA individuals with MMSE scores < 27, who had not been diagnosed with either MCI or AD, produced little difference in the atrophy estimates other than the significant interaction between FreeSurfer and the OA individuals becoming a non-significant trend (Supplementary Tables S4, S5).

Laterality Analysis

The right hippocampus was larger than the left hippocampus across all age groups, times and segmentation methods. FreeSurfer and manual tracing produced similar levels of laterality at Time-1. At Time-2 laterality increased for manual tracing and decreased for FreeSurfer (Table 5).

Similarity Analysis

The voxel by voxel comparison of manually traced hippocampi to FreeSurfer segmented hippocampi was performed for the baseline scans of the MA age group and the

Table 3 Mean annualised atrophy rates for manual tracing and FreeSurfer

	n	Left (SD)	95% CI	Right (SD)	95% CI	Total (SD)	95% CI
Atrophy mm^3							
MA							
Manual	230	−14.60 (34.54)	−19.08, −10.11	1.11 (34.94)	−3.43, 5.65	−13.48 (60.87)	−21.39, −5.57
FreeSurfer	230	−7.030 (29.05)	−10.80, −3.25	−9.44 (27.58)	−13.02, −5.85	−16.46 (49.74)	−22.93, −10.00
OA							
Manual	190	−15.33 (42.16)	−21.37, −9.30	−6.57 (46.80)	−13.27, 0.12	−21.91 (80.65)	−33.45, −10.36
FreeSurfer	190	−13.40 (30.34)	−17.74, −9.06	−27.43 (30.82)	−31.84, −23.02	−40.83 (51.92)	−48.26, −33.40
Atrophy %							
MA							
Manual	230	−0.39 (1.00)	−0.52, −0.25	0.07 (1.01)	−0.06, 0.20	−0.17 (0.88)	−0.28, −0.05
FreeSurfer	230	−0.15 (0.69)	−0.24, −0.06	−0.20 (0.63)	−0.29, −0.12	−0.18 (0.57)	−0.26, −0.11
OA							
Manual	190	−0.43 (1.32)	−0.62, −0.24	−0.15 (1.41)	−0.35, 0.05	−0.30 (1.23)	−0.48, −0.12
FreeSurfer	190	−0.34 (0.79)	−0.45, −0.22	−0.68 (0.78)	−0.79, −0.57	−0.51 (0.65)	−0.61, −0.42

MA middle age, *OA* older age, *95% CI* 95% confidence interval

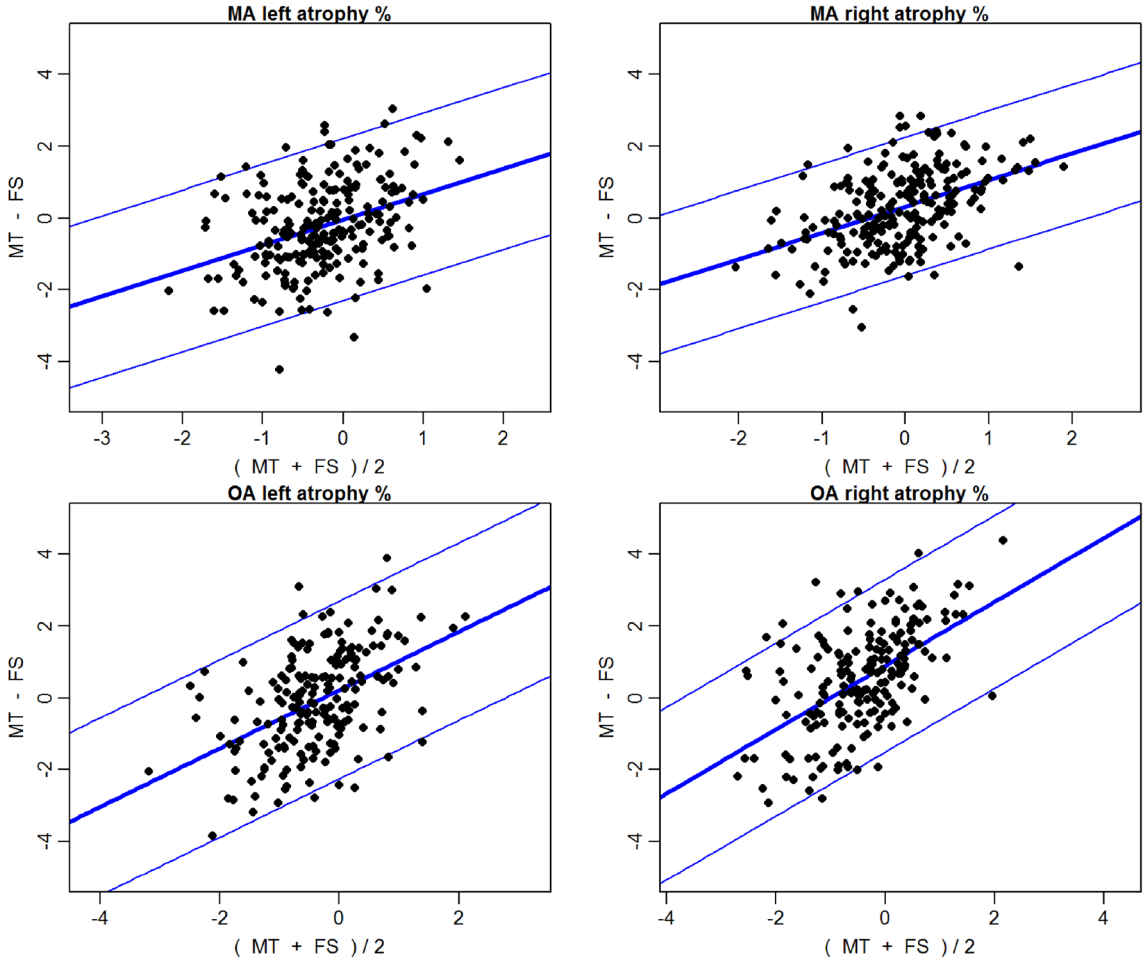

Fig. 1 Bland–Altman (BA) mean difference plots comparing manual tracing (MT) annualised atrophy rates to FreeSurfer (FS) atrophy rates for the middle age (MA) group (top row) and the older age (OA) group (bottom row). The thick lines on each plot show the mean bias and the limits of agreement is indicated by the thinner lines at ±2SD of the mean difference. All plots demonstrate a consistent pattern where the absolute difference between manual tracing and FreeSurfer increases as the absolute value of the mean estimate increases. This means that the gap between the estimates for FreeSurfer and Manual tracing will increase as the rate of atrophy increases

Table 4 Predictors of annualised percentage atrophy over 8 years

Predictors	B	95% CI	SE	T	p
(Intercept)	−0.3946	−0.5294 −0.2597	0.0688	−5.734	<0.0001
FreeSurfer	0.2617	0.1045, 0.4190	0.0802	3.263	0.0011
Age group (OA = 1)	−0.1250	-0.2585, 0.0084	0.0681	−1.836	0.0665
Male	−0.0266	−0.1210, 0.0678	0.0482	−0.5521	0.5810
Side (right = 1)	0.3732	0.2409, 0.5055	0.0675	5.530	<0.0001
Age (centred)	−0.0330	−0.0666, 0.0007	0.0172	−1.921	0.0549
FreeSurfer*OA	−0.1983	−0.3862, −0.0104	0.0959	−2.068	0.0388
FreeSurfer*Right	−0.5582	-0.7452, −0.3711	0.0954	−5.849	<0.0001
	Mc Fadden R2 = 0.0428				

95% CI 95% confidence interval around the B coefficient, *MA* middle age, *OA* older age, *Side* left or right hippocampus, *Age* age centred within age group, Reference method is manual tracing

Table 5 Comparison of laterality index between manual tracing and FreeSurfer

Age Group	Method	Time	N	LI (SD)	95% CI
MA	Manual	Time-1	238	−0.97 (3.58)	(−1.43, −0.52)
		Time-2	238	−2.95 (3.79)	(−3.43, −2.47)
	FreeSurfer	Time-1	238	−0.87 (3.09)	(−1.26, −0.47)
		Time-2	238	−0.65 (3.10)	(−1.05, −0.25)
OA	Manual	Time-1	194	−1.98 (4.19)	(−2.57, −1.39)
		Time-2	194	−3.15 (4.87)	(−3.84, −2.46)
	FreeSurfer	Time-1	194	−2.03 (3.23)	(−2.49, −1.57)
		Time-2	194	−0.65 (4.08)	(−1.23, −0.07)

MA middle age, *OA* older age, *LI* laterality index, *95% CI* 95% confidence interval

OA group to enable the calculation of Dice similarity coefficients. Pairs with low Dice coefficients (<0.6) were visually inspected and nine cases with registration errors and six cases with visually obvious FreeSurfer segmentation errors were identified and excluded from the Dice overlap coefficients shown in Table 6.

GLM models to investigate the impact of age, age group, gender, ICV, and laterality on Dice coefficients indicated a significant main effect for laterality where there was greater overlap for the right hippocampus (Table 7). Post hoc analysis revealed a significant positive relationship between Dice coefficients and raw traced hippocampal volume. The increase in Dice coefficients per cm^3 for the MA group is 0.032 (CI 0.021, 0.042) left and 0.019 (CI 0.009, 0.030) right. For the OA group, the increase was 0.045 (CI 0.035, 0.054) left and 0.028 (CI 0.019, 0.037) right (Supplementary Fig. S2). The relationship increased when the analysis was repeated using normalised hippocampal volumes indicating the effect was unrelated to ICV.

A visual inspection of the merged segmentations, with one slice examined from each orientation, indicated that there was good agreement between the methods for voxels with low intensity that were clearly CSF. The additional hippocampal volume attributed to FreeSurfer was mainly related to the inclusion of additional voxels in the lateral and dorsal surfaces for most of the length of the hippocampus. Whereas a smaller number of additional voxels included by manual tracing tended to be on the medial and ventral surfaces (Fig. 2a, b, Fig. 3).

A more detailed inspection of the segmentations with the highest dice coefficients in each sample identified additional differences between manual tracing and FreeSurfer. Firstly, in manual tracing the border of the hippocampus with the amygdala follows and includes the line of the alveus (composed of white matter) whereas in FreeSurfer the border appears slightly superior to the alveus (Fig. 2c, d). Secondly, FreeSurfer included voxels that contain a mixture of tissues to a greater extent than manual tracing consistent

Table 6 Similarity of segmented scans measured by Dice coefficient

	Dice coefficient				
	Pairs	Left (SD)	95% CI	Right (SD)	95% CI
MA	226	0.7720 (0.0388)	(0.7670, 0.7771)	0.7803 (0.0327)	(0.7760, 0.7845)
OA	191	0.7754 (0.0332)	(0.7707, 0.7801)	0.7840 (0.0297)	(0.7798, 0.7883)

The Dice coefficient measures the proportion of overlapping voxels between manual tracing and FreeSurfer segmentation

Pairs number matched segmentations, *95% CI* 95% confidence interval, *MA* middle age, *OA* older age

Table 7 Predictors of similarity between manual tracing and FreeSurfer automated segmentation of the hippocampus measured by Dice overlap coefficients

Predictors	B	95% CI	SE	T	p
(Intercept)	0.7648	0.7405, 0.7890	0.0124	61.810	<0.0001
Age group (OA = 1)	0.0036	−0.0011, 0.0083	0.0024	1.515	0.1301
Male	−0.0058	−0.0116, −0.0001	0.0030	−1.923	0.0548
Side (right = 1)	0.0084	0.0038, 0.0130	0.0023	3.578	0.0004
Age (centred)	0.0016	−0.0001, 0.0033	0.0008	1.898	0.0581
ICV	0.0000	0.0000, 0.0000	0.0000	0.990	0.3223
	Mc Fadden R2 = 0.0263				

95% CI 95% confidence interval around B coefficient, *MA* middle age, *OA* older age, *Side* left or right hippocampus, *Age (centred)* age centred within group, *ICV* estimated total intracranial volume

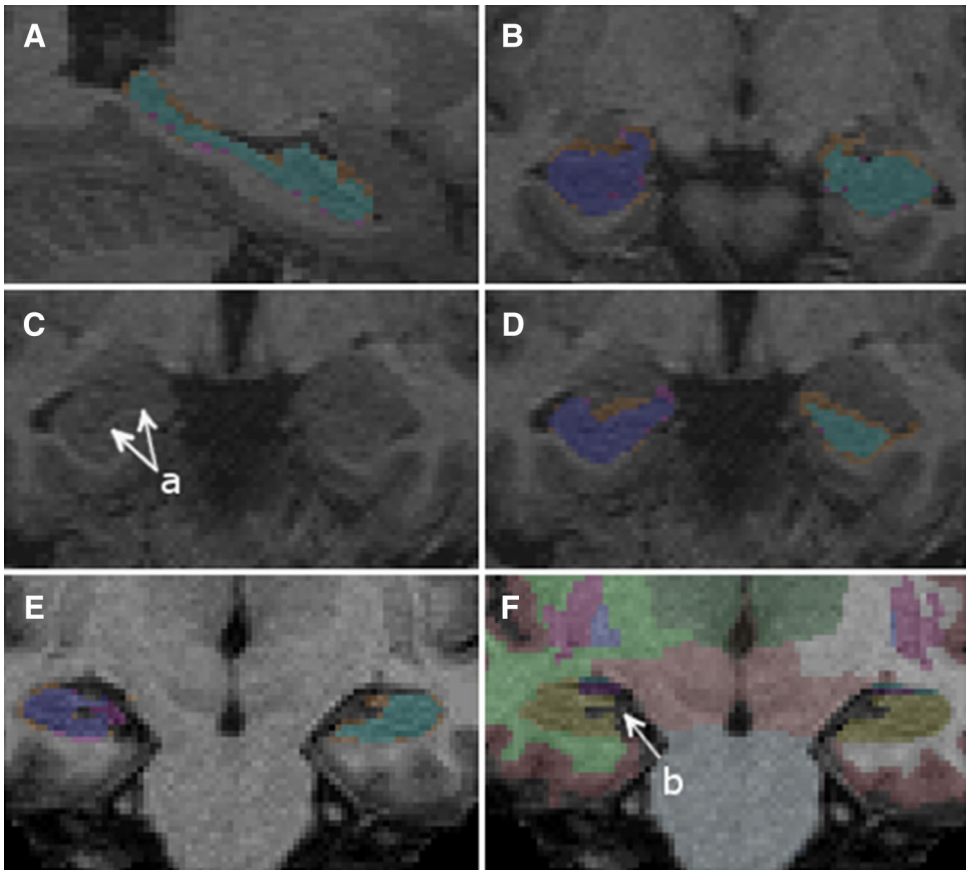

Fig. 2 Differences between manual tracing and FreeSurfer segmentations. **a** Sagittal view of the left hippocampus with additional FreeSurfer only voxels (Orange) mainly along the dorsal surfaces and additional manual tracing voxels (Pink) mainly along ventral surfaces. **b** Coronal view of the scan in Panel **a**, showing right hippocampus overlapping voxels (Blue), left hippocampus overlapping voxels (Aqua), FreeSurfer only voxels (Orange) and manual tracing only voxels (Pink). **c** and **d** The left panel **c** shows the alveus indicated by the arrows (a) and the right panel **d** shows the same scan with the merged segmentation overlayed. The FreeSurfer superior border includes additional voxels (Orange) superior to the alveus in this scan. **e** and **f** The left panel **e** shows the merged manual tracing and FreeSurfer segmentations. The area indicated by arrow (b) in the right hand panel **f** are voxels that were included as hippocampus by manual tracing but not labelled to any structure by FreeSurfer. The colour scheme for panels **a–e, g, h** shows the overlap between manual tracing and FreeSurfer; FreeSurfer; Blue = Right hippocampus overlap (TP); Orange = FreeSurfer only (FP), Pink = Manual tracing only (FN), Aqua = Left hippocampus overlap (TP). The colour scheme for panel F is the standard FreeSurfer label colours; Brown = hippocampus, Purple = Inferior lateral ventricle.*TP* true positive, *FP* false positive, *FN* false negative

with previous studies (Cherbuin et al. 2009). Third, small cavities of CSF (Maller et al. 2011) enclosed within the hippocampus were sometimes labelled as hippocampal tissue by FreeSurfer. Finally, there were a small number of cases where some of the medial voxels were not labelled to any structure by FreeSurfer (Fig. 2e, f).

Discussion

We compared manual tracing to FreeSurfer segmentation (Reuter et al. 2012) in healthy middle age and older age individuals over a period of 8 years. The main findings were that FreeSurfer produced higher volumes than manual tracing. Secondly, there were laterality differences in the amount of overlap between the segmentations produced by manual tracing and FreeSurfer. Third, both methods produced atrophy rates that were somewhat lower than previously reported estimates. Fourth, both methods produced similar total rates for middle aged individuals and for older individuals, FreeSurfer produced significantly higher estimates. Finally changes in laterality over time, where laterality increased for manual tracing and decreased for FreeSurfer, led to laterality differences in atrophy rates.

At baseline, the FreeSurfer volumes were greater than those produced by manual tracing in line with most previous comparison studies (Arnold et al. 2015; Cherbuin et al. 2009; Doring et al. 2011; Morey et al. 2008; Shen et al. 2010; Tae et al. 2008; Wenger et al. 2014). Similar to Cherbuin et al. (2009), inspection of the manual and

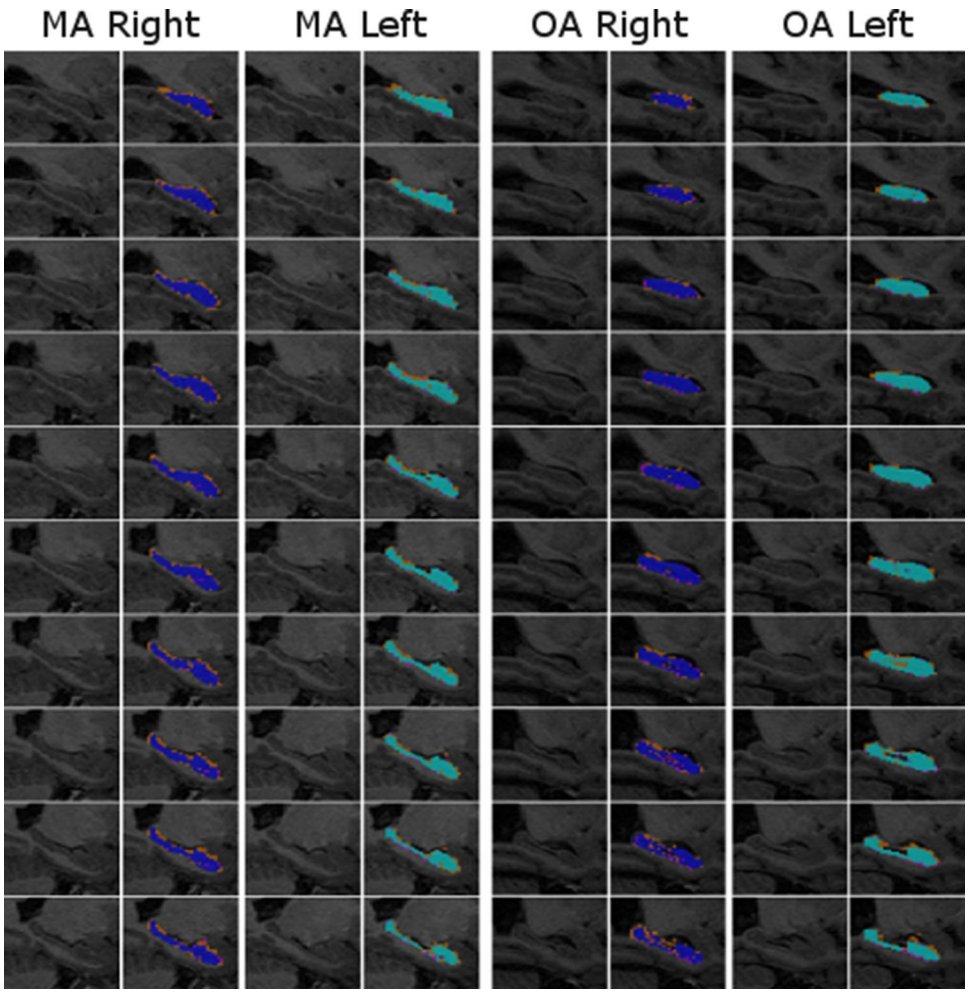

Fig. 3 Slice by slice sagittal views of the overlap between FreeSurfer and manual tracing from two participant scans. There is good agreement between the methods for voxels with low intensity that are clearly CSF. The additional hippocampal volume attributed to FreeSurfer was predominantly related to the inclusion of additional voxels in the lateral and dorsal surfaces for most of the length of the hippocampus. Whereas a smaller number of additional voxels included by manual tracing tended to be on the medial and ventral surfaces, or related to cavities within the hippocampi. Colour scheme for the overlap between manual tracing and FreeSurfer; Blue = Right hippocampus overlap (TP); Orange = FreeSurfer only (FP), Pink = Manual tracing only (FN), Aqua = Left hippocampus overlap (TP). *TP* true positive, *FP* false positive, *FN* false negative

FreeSurfer segmentations suggests that the differences are due to three main factors. Firstly, the inclusion by FreeSurfer of boundary voxels along the lateral and dorsal surfaces that contained a mixture of grey and, white matter or CSF. Secondly, the inclusion by FreeSurfer of small cavities of CSF (Maller et al. 2011) as hippocampal tissue. Third, small differences in the location of the border between the hippocampus and the amygdala.

We hypothesised that the difference between manual tracing and FreeSurfer segmentations, in terms of proportion of overlap, would be less pronounced in the older sample as a consequence of higher atrophy in older individuals increasing the amount of high contrast hippocampal boundaries. Absolute volume differences were slightly smaller for the older aged group compared to the middle aged group, as would be expected given the smaller hippocampal volumes in the older group (Shen et al. 2010; Wenger et al. 2014). We also found that there was good agreement between the methods for voxels with low intensity that were clearly CSF (Figs. 2, 3). Yet while the proportion of overlap was slightly higher for the older aged group, the association was not significant.

However, there was a greater proportion of overlap (similarity) between the right hippocampal regions defined by manual tracing and FreeSurfer. One possible explanation is the presence of a visual laterality bias in manual tracing where objects traced on the left side of the screen produce larger volumes (Maltbie et al. 2012; Rogers et al. 2012; Wenger et al. 2014). In the current study, the manual tracing was always performed on images presented in radiological orientation with the right hippocampus on the left side of the screen. When interpreting laterality differences in studies involving manual tracing, the potential for a visual bias needs to be considered when the raters have not been blinded to laterality and the direction of the bias would depend upon the display orientation (Radiological or Neurological). Ideally, raters should be blinded to laterality and the display orientation should be randomised. Alternatively, FreeSurfer may be generating proportionally smaller volumes for the right hippocampus. The FreeSurfer default atlas used in

this study is based on 39 brains (aged 18–87) manually segmented in accordance with the conventions of the Center for Morphometric Analysis standard (CMA;Fischl et al. 2002, 2004; Makris et al. 2004; Reuter et al. 2012; Sabuncu et al. 2010). The CMA standard indicates radiological orientation, but does not indicate which side is traced first. The CMA protocol also indicates that the hippocampi are traced after all surrounding structures have been traced. So, while the FreeSurfer atlas is based on a manual tracing protocol that might positively bias right hippocampal volumes, the bias may be limited by the hippocampi being traced last in the atlas, and this could also explain why automated segmentation was more similar to manual tracing for the right hippocampus in the current study. It is also possible that the increased similarity could be partially related to the right hippocampus being larger than the left.

Indeed, post hoc analysis revealed that the proportion of overlap increased as hippocampal size increased. FreeSurfer includes additional boundary voxels (Fig. 2; Cherbuin et al. 2009) resulting in a slightly inflated volume compared to manual tracing and these additional voxels will have a proportionally smaller impact on larger hippocampi. The effect is unrelated to overall head size, so it does not create an accuracy bias related to gender. However in studies that investigate laterality, the bias related to hippocampal size may need to be considered.

There were a small number of FreeSurfer scans that had visually obvious segmentation errors that were not identified by the segmentation checks as outliers of hippocampal, total grey or total white matter volumes but were identified when comparing their segmentation results to manual tracings. Including these minor segmentation errors would have little effect in the current study given the large sample size. However in smaller samples, such segmentation errors could have a meaningful impact. Thus, when FreeSurfer is used for small samples, all scans should be visually checked to exclude or correct minor segmentation errors.

Total atrophy rates for both manual tracing and FreeSurfer in this study are lower, but within the confidence interval range of those reported in the meta-analyses by Fraser et al. (2015) and Barnes et al. (2009a). However, the total atrophy rates were similar to a recent large cross-sectional population study (Schmidt et al. 2018). The left and right atrophy estimates for both methods were within the confidence interval ranges of Fraser et al. (2015). The lower total atrophy rates could be related to the characteristics of the population samples used in the current study. Firstly, the samples in the current study are relatively large and have a narrow age range of ± 2 years that likely results in more precise atrophy estimates and narrower confidence intervals. Secondly, participants were relatively healthy given that they remained in the PATH study for at least 8 years and met the stringent exclusion criteria. Third, the individuals in the older age PATH cohort were somewhat younger than controls in AD focused datasets such as ADNI.

Total FreeSurfer atrophy rates were similar to those for manual tracing in middle age individuals and higher for older age individuals in contrast to the meta-analysis by Fraser et al. (2015) which found that manual tracing produced higher rates of atrophy than automated methods. Fraser et al. (2015) suggested that the difference could be the result of automated segmentation including non-hippocampal tissues with lower atrophy rates, or that some of the studies used manual tracing protocols that excluded the hippocampal tail (e.g. Watson et al. 1997) combined with differences in atrophy rates between the hippocampal tail and the rest of the hippocampus. The results of our study did not provide support for either alternatives as the larger FreeSurfer segmented volumes did not result in lower total atrophy rates, and the manual atrophy rates did not change significantly when the hippocampal tail was excluded as part of the sensitivity analysis. However, it is possible that the higher manual atrophy rates from the meta-analysis could have been the result of the manual tracing studies having small sample sizes (median $n = 26$) resulting in low power leading to overestimated effect sizes (Button et al. 2013).

There were some other areas where the methods differed. The most significant was that manual tracing atrophy rates were higher for the left hippocampus than the right hippocampus consistent with the meta-analyses of Shi et al. (2009) and Fraser et al. (2015), whereas FreeSurfer atrophy estimates for the right hippocampus were higher than those for the left hippocampus. Secondly, the standard deviations of the FreeSurfer atrophy estimates were consistently smaller than those of manual tracing, resulting in more precise estimates with narrower confidence intervals, suggesting that there may be less error in the FreeSurfer estimates. This result is consistent with the finding from a previous longitudinal comparison by Mulder et al. (2014) that FreeSurfer had lower variance and higher test–retest reliability compared to manual tracing.

Limitations

We acknowledge some limitations with the current study. Manually traced ICVs were not available for normalising the manually traced hippocampal volumes and the ICV generated by the FreeSurfer cross-sectional pipeline were used instead. While not optimal, it is not uncommon to use ICVs generated by automated methods to normalise manual tracing volumes and it is unlikely it would have had a meaningful effect on the present findings.

The manual tracing protocol used in the study was selected in 1999 at the commencement of the PATH population study and there may be some differences to the current

gold standard in manual tracing (Boccardi et al. 2015; Frisoni et al. 2015). When using manual tracing for long running studies the choice is to freeze the tracing protocol to the start of the study or risk introducing confounds by incorporating technique improvements over time as it is not realistic to manually re-trace earlier time points in large studies. An advantage that automated methods such as FreeSurfer have over manual tracing for long running population studies is that it is possible to reprocess all time points to incorporate new technique improvements as they are developed. For example, the latest version of FreeSurfer (version 6.0) includes an additional processing stream that segments hippocampi into subfields that results in lower overall hippocampal volumes (Schmidt et al. 2018).

Conclusions

In conclusion, we found that FreeSurfer produced higher volumes than manual tracing. The proportion of overlap between the hippocampal structures created by manual tracing and FreeSurfer segmentation was higher for the right hippocampus. Over time, laterality increased for manual tracing and decreased for FreeSurfer leading to laterality differences in left and right estimated atrophy rates. The total hippocampal atrophy rates from both methods increased with age. FreeSurfer total hippocampal atrophy rates were similar to manual tracing estimates for middle aged individuals and significantly higher for older age individuals. FreeSurfer reliably segments the hippocampus producing atrophy rates that are comparable to manual tracing with some biases that need to be considered in study design. FreeSurfer is suited for use in large longitudinal studies where it is not cost effective to use manual tracing.

Acknowledgements The authors are grateful to Chantal Réglade-Meslin, Jerome Maller, Peter Butterworth, Simon Easteal, Helen Christensen, Patricia Jacomb, Karen Maxwell, and the PATH interviewers. The study was supported by an Australian Government Research Training Program (RTP) Scholarship, National Health and Medical Research Council (NHMRC) Grant Nos. 973302, 179805,350833 157125, and Australian Research Council (ARC) Grant No. 130101705. Kaarin Anstey was funded by NHMRC Fellowship No.1002560. This research was partly undertaken on the National Computational Infrastructure (NCI) facility in Canberra, Australia, which is supported by the Australian Commonwealth Government. The authors declare no competing financial interests. This research is supported by an Australian Government Research Training Program (RTP) Scholarship. This study is NOT industry sponsored.

Author Contributions MAF contributed to the design of the study, conducted all statistical analyses, and managed all aspects of manuscript preparation and submission. MES contributed to the design of the study and the statistical analyses, provided methodological input and theoretical expertise, and contributed to writing and editing of the manuscript. KJA contributed to the design of the study, provided methodological input and theoretical expertise, and contributed to writing and editing of the manuscript. NC contributed to the design of the study and the statistical analyses, provided methodological input and theoretical expertise, and contributed to writing and editing of the manuscript.

Compliance with Ethical Standards

Conflict of interest The authors have reported no conflicts of interest.

References

Anstey KJ et al (2012) Cohort profile: the PATH through life project. Int J Epidemiol 41:951–960. https://doi.org/10.1093/ije/dyr025

Arnold SJ et al (2015) Hippocampal volume is reduced in schizophrenia and schizoaffective disorder but not in psychotic bipolar I disorder demonstrated by both manual tracing and automated parcellation (FreeSurfer). Schizophr Bull 41:233–249. https://doi.org/10.1093/schbul/sbu009

Barnes J et al (2004) Differentiating AD from aging using semiautomated measurement of hippocampal atrophy rates Neuroimage https://doi.org/10.1016/j.neuroimage.2004.06.028

Barnes J et al (2008) A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus NeuroImage 40:1655–1671 https://doi.org/10.1016/j.neuroimage.2008.01.012

Barnes J et al (2009a) A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. Neurobiol Aging 30:1711–1723. https://doi.org/10.1016/j.neurobiolaging.2008.01.010

Barnes J, Ourselin S, Fox NC (2009b) Clinical application of measurement of hippocampal atrophy in degenerative dementias. Hippocampus 19:510–516. https://doi.org/10.1002/hipo.20617

Bergouignan L et al (2008) Can voxel based morphometry, manual segmentation and automated segmentation equally detect hippocampal volume differences in acute depression? Neuroimage 45:29–37. https://doi.org/10.1016/j.neuroimage.2008.11.006

Bland JM, Altman DG (1999) Measuring agreement in method comparison studies Stat Methods Med Res 8:135–160

Boccardi M et al (2011) Survey of protocols for the manual segmentation of the Hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J Alzheimer's Dis 26:61–75. https://doi.org/10.3233/JAD-2011-0004

Boccardi M et al (2015) Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance. Alzheimer's Dement 11:126–138. https://doi.org/10.1016/j.jalz.2014.02.009

Braak H, Braak E, Bohl J, Reintjes R (1996) Age, neurofibrillary changes, A beta-amyloid and the onset of Alzheimer's disease. Neurosci Lett 210:87–90

Buckner RL, Head D, Parker J, Fotenos AF, Marcus D, Morris JC, Snyder AZ (2004) A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. Neuroimage 23:724–738. https://doi.org/10.1016/j.neuroimage.2004.06.018

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376. https://doi.org/10.1038/nrn3475

Carstensen B (2010) Comparing methods of measurement: extending the LoA by regression. Stat Med 29:401–410. https://doi.org/10.1002/sim.3769

Cash DM et al (2015) Assessing atrophy measurement techniques in dementia: results from the MIRIAD atrophy challenge

NeuroImage 123:149–164 https://doi.org/10.1016/j.neuroimage.2015.07.087
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 https://doi.org/10.1371/journal.pone.0005265
Cherbuin N, Sargent-Cox K, Easteal S, Sachdev P, Anstey KJ (2014) Hippocampal atrophy is associated with subjective memory decline: the PATH through life study. Am J Geriatr Psychiatry 23(5):446–455
de Flores R et al (2014) Effects of age and Alzheimer's disease on hippocampal subfields: comparison between manual and freesurfer volumetry. Hum Brain Mapp https://doi.org/10.1002/hbm.22640
Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302. https://doi.org/10.2307/1932409
Dill V, Franco AR, Pinho MSS (2015) Automated methods for hippocampus segmentation: the evolution and a review of the state of the art. Neuroinformatics 13:133–150. https://doi.org/10.1007/s12021-014-9243-4
Doring TM, Kubo TT, Cruz LC Jr, Juruena MF, Fainberg J, Domingues RC, Gasparetto EL (2011) Evaluation of hippocampal volume based on MR imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. J Magn Reson Imaging 33:565–572. https://doi.org/10.1002/jmri.22473
Draganski B, Gaser C, Kempermann G (2006) Temporal and spatial dynamics of brain structure changes during extensive learning. J Neurosci https://doi.org/10.1523/JNEUROSCI.4628-05.2006
Erickson KI et al (2011) Exercise training increases size of hippocampus and improves memory. Proc Natl Acad Sci USA 108:3017–3022. https://doi.org/10.1073/pnas.1015950108
Fischl B et al (2002) Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33:341–355. https://doi.org/10.1016/S0896-6273(02)00569-X
Fischl B et al (2004) Automatically parcellating the human cerebral cortex. Cereb Cortex 14:11–22. https://doi.org/10.1093/cercor/bhg087
Fjell AM et al (2013) Critical ages in the life course of the adult brain: nonlinear subcortical aging. Neurobiol Aging 34:2239–2247. https://doi.org/10.1016/j.neurobiolaging.2013.04.006
Fraser MA, Shaw ME, Cherbuin N (2015) A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. Neuroimage https://doi.org/10.1016/j.neuroimage.2015.03.035
Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6:67–77
Frisoni GB et al (2015) The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimer's Dementia 11:111–125. https://doi.org/10.1016/j.jalz.2014.05.1756
Gronenschild EHBM, Habets P, Jacobs HIL, Mengelers R, Rozendaal N, van Os J, Marcelis M (2012) The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS ONE https://doi.org/10.1371/journal.pone.0038234
Hsu YY, Schuff N, Du AT, Mark K, Zhu X, Hardin D, Weiner MW (2002) Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. J Magn Reson Imaging 16:305–310. https://doi.org/10.1002/jmri.10163
Jinno S (2015) Aging affects new cell production in the adult hippocampus: a quantitative anatomic review. J chem Neuroanat https://doi.org/10.1016/j.jchemneu.2015.10.009
Leung KK et al (2010) Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 51:1345–1359. https://doi.org/10.1016/j.neuroimage.2010.03.018
Makris N et al (2004) General brain segmentation: method and utilization Massachusetts General Hospital Boston, MA, USA
Maller JJ, Réglade-Meslin C, Anstey KJ, Sachdev P (2006) Sex and symmetry differences in hippocampal volumetrics: before and beyond the opening of the crus of the fornix. Hippocampus 16:80–90. https://doi.org/10.1002/hipo.20133
Maller JJ et al (2011) Hippocampal sulcal cavities: prevalence, risk factors and relationship to memory impairment. Brain Res 1368:222–230. https://doi.org/10.1016/j.brainres.2010.10.089
Maltbie E et al (2012) Asymmetric bias in user guided segmentations of. brain structures Neuroimage 59:1315–1323. https://doi.org/10.1016/j.neuroimage.2011.08.025
Morey RA et al (2008) A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes Neuroimage 45:855–866. https://doi.org/10.1016/j.neuroimage.2008.12.033
Mulder ER et al (2014) Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 92:169–181. https://doi.org/10.1016/j.neuroimage.2014.01.058
Pfefferbaum A, Sullivan EV (2015) Cross-sectional versus longitudinal estimates of age-related changes in the adult brain: overlaps and discrepancies. Neurobiol Aging 36:2563–2567. https://doi.org/10.1016/j.neurobiolaging.2015.05.005
Raz N, Ghisletta P, Rodrigue KM, Kennedy KM, Lindenberger U (2010) Trajectories of brain aging in middle-aged and older adults: regional and individual differences. Neuroimage 51:501–511. https://doi.org/10.1016/j.neuroimage.2010.03.020
Reuter M, Schmansky NJ, Rosas DH, Fischl B (2012) Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61:1402–1418. https://doi.org/10.1016/j.neuroimage.2012.02.084
Rogers BP, Sheffield JM, Luksik AS, Heckers S (2012) Systematic error in Hippocampal volume asymmetry measurement is minimal with a manual segmentation protocol. Front Neurosci 6:179. https://doi.org/10.3389/fnins.2012.00179
Sabuncu MR, Yeo TBT, Leemput VK, Fischl B, Golland P (2010) A generative model for image segmentation based on label fusion. IEEE Trans Med Imaging 29:1714–1729. https://doi.org/10.1109/tmi.2010.2050897
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 181:219–225. https://doi.org/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 registered magnetic resonance imaging. Arch Neurol 60:989–994. https://doi.org/10.1001/archneur.60.7.989
Schmidt MF, Storrs JM, Freeman KB, Jack CR, Turner ST, Griswold ME, Mosley TH (2018) A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan. Hum Brain Mapp https://doi.org/10.1002/hbm.24017
Shaw ME, Abhayaratna WP, Sachdev PS, Anstey KJ, Cherbuin N (2016a) Cortical thinning at midlife: the PATH through life study. Brain Topogr 29:875–884. https://doi.org/10.1007/s10548-016-0509-z
Shaw ME, Sachdev PS, Anstey KJ, Cherbuin N (2016b) Age-related cortical thinning in cognitively healthy individuals in their 60s: the PATH through life study. Neurobiol Aging 39:202–209. https://doi.org/10.1016/j.neurobiolaging.2015.12.009
Shaw ME, Nettersheim J, Sachdev PS, Anstey KJ, Cherbuin N (2017) Higher fasting plasma glucose is associated with increased

cortical thinning over 12 years: the PATH through life study. Brain Topogr 30:408–416. https://doi.org/10.1007/s10548-017-0544-4

Shen L et al (2010) Comparison of manual and automated determination of Hippocampal volumes in MCI and early AD. Brain Imaging Behav 4:86–95. https://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. https://doi.org/10.1002/hipo.20573

Sled JG, Zijdenbos AP, Evans AC (1998) A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17:87–97. https://doi.org/10.1109/42.668698

Tae W, Kim S, Lee K, Nam E-C, Kim K (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. https://doi.org/10.1007/s00234-008-0383-9

Walsh EI, Shaw M, Sachdev P, Anstey KJ, Cherbuin N (2017) Brain atrophy in ageing: Estimating effects of blood glucose levels vs. other type 2 diabetes effects. Diabetes Metab https://doi.org/10.1016/j.diabet.2017.06.004

Watson C, Jack CR, Cendes F (1997) Volumetric magnetic resonance imaging. Clinical applications and contributions to the understanding of temporal lobe epilepsy. Arch Neurol 54:1521–1531

Wenger E et al (2014) Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains. Hum Brain Mapp 35:4236–4248. https://doi.org/10.1002/hbm.22473

Woollett K, Maguire EA (2011) Acquiring "the Knowledge" of London's layout drives structural brain changes. Curr Biol 21(24):2109–2114