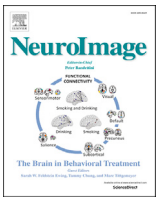
NeuroImage 210 (2020) 116563

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



Test-retest reliability of FreeSurfer automated hippocampal subfield segmentation within and across scanners

Emma M. Brown [a,b,*], Meghan E. Pierce [b,c], Dustin C. Clark [a], Bruce R. Fischl [d,g,i], Juan E. Iglesias [d,e,i], William P. Milberg [b,c,h], Regina E. McGlinchey [b,c,h], David H. Salat [a,b,d,f,g]

[a] Neuroimaging Research for Veterans (NeRVe) Center, VA Boston Healthcare System, Boston, MA, USA
[b] Translational Research Center for TBI and Stress Disorders (TRACTS), VA Boston Healthcare System, Boston, MA, USA
[c] Department of Psychiatry, Harvard Medical School, Boston, MA, USA
[d] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA
[e] Centre of Medical Image Computing (CMIC), Department of Medical Physics and Biomedical Engineering, University College London (UCL), London, UK
[f] Brain Aging and Dementia (BAnD) Laboratory, A. A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA
[g] Department of Radiology, Harvard Medical School, Boston, MA, USA
[h] Geriatric Research, Education and Clinical Center (GRECC), VA Boston Healthcare System, Boston, MA, USA
[i] Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

ARTICLE INFO

Keywords:
Test-retest reliability
FreeSurfer
Hippocampus
Hippocampal subfields
Magnetic resonance imaging
Longitudinal

ABSTRACT

The human hippocampus is vulnerable to a range of degenerative conditions and as such, accurate *in vivo* measurement of the hippocampus and hippocampal substructures via neuroimaging is of great interest for understanding mechanisms of disease as well as for use as a biomarker in clinical trials of novel therapeutics. Although total hippocampal volume can be measured relatively reliably, it is critical to understand how this reliability is affected by acquisition on different scanners, as multiple scanning platforms would likely be utilized in large-scale clinical trials. This is particularly true for hippocampal subregional measurements, which have only relatively recently been measurable through common image processing platforms such as FreeSurfer. Accurate segmentation of these subregions is challenging due to their small size, magnetic resonance imaging (MRI) signal loss in medial temporal regions of the brain, and lack of contrast for delineation from standard neuroimaging procedures.

Here, we assess the test-retest reliability of the FreeSurfer automated hippocampal subfield segmentation procedure using two Siemens model scanners (a Siemens Trio and Prisma^{fit} Trio upgrade). T1-weighted images were acquired for 11 generally healthy younger participants (two scans on the Trio and one scan on the Prisma^{fit}). Each scan was processed through the standard cross-sectional stream and the recently released longitudinal pipeline in FreeSurfer v6.0 for hippocampal segmentation. Test-retest reliability of the volumetric measures was examined for individual subfields as well as percent volume difference and Dice overlap among scans and intraclass correlation coefficients (ICC). Reliability was high in the molecular layer, dentate gyrus, and whole hippocampus with the inclusion of three time points with mean volume differences among scans less than 3%, overlap greater than 80%, and ICC >0.95. The parasubiculum and hippocampal fissure showed the least improvement in reliability with mean volume difference greater than 5%, overlap less than 70%, and ICC scores ranging from 0.78 to 0.89. Other subregions, including the CA regions, were stable in their mean volume difference and overlap (<5% difference and >75% respectively) and showed improvement in reliability with the inclusion of three scans (ICC > 0.9). Reliability was generally higher within scanner (Trio-Trio), however, Trio-Prisma^{fit} reliability was also high and did not exhibit an obvious bias. These results suggest that the FreeSurfer automated segmentation procedure is a reliable method to measure total as well as hippocampal subregional volumes and may be useful in clinical applications including as an endpoint for future clinical trials of conditions affecting the hippocampus.

* Corresponding author. Neuroimaging Research for Veterans (NeRVe) Center, VA Boston Healthcare System, USA.
E-mail address: emma.brown2@va.gov (E.M. Brown).

https://doi.org/10.1016/j.neuroimage.2020.116563
Received 1 August 2019; Received in revised form 13 January 2020; Accepted 15 January 2020
Available online 21 January 2020

1. Introduction

The hippocampus supports a wide variety of cognitive functions, including memory, reward processing, and executive processes (Fortin et al., 2002; Scoville and Milner, 1957; Winocur et al., 2006). This limbic region is extremely complex anatomically at the histological level, yet has major subregions with distinct functional properties and vulnerability to pathologies that can be identified via brain imaging procedures (Brickman et al., 2011; Hoge and Kesner, 2007; Langston et al., 2010; Small et al., 2004). These subregions are variably defined yet commonly include the dentate gyrus, subiculum, parasubiculum, entorhinal cortex, and the four cornu ammonis (CA) regions (Benarroch, 2013; Horovitz and Richter-Levin, 2015). Previous studies have demonstrated high plasticity in the hippocampus and its subfields across the lifespan (Benarroch, 2013; Ergorul and Eichenbaum, 2004; Fortin et al., 2002; Horovitz and Richter-Levin, 2015; Leuner and Gould, 2010; Mcewen, 1999) in relation to episodic memory, undergoing a continual process of strengthening, weakening, and altering (Horner and Doeller, 2017). This region is also sensitive to a range of negative influences on neural integrity including hypoxia, stress hormones, and Alzheimer's disease (AD) pathology (Brickman et al., 2011; de Flores et al., 2015; Di Paola et al., 2008; Mcewen, 1999; Teicher et al., 2012). Thus, alterations in the hippocampus or its subfields have been reported across several neuropsychiatric and degenerative disorders, including Schizophrenia (SZD), posttraumatic stress disorders (PTSD), and Alzheimer's disease (de Flores et al., 2015; La Joie et al., 2013; Leuner and Gould, 2010; Scoville and Milner, 1957; Teicher et al., 2012; Winocur et al., 2006). Given the critical role of the hippocampus in cognitive health and vulnerability to disease (Brickman et al., 2011; Ergorul and Eichenbaum, 2004; Fortin et al., 2002; Hoge and Kesner, 2007; Horner and Doeller, 2017; Leuner and Gould, 2010; Scoville and Milner, 1957; Small et al., 2004), reliable measurement of these substructures is a major goal of neuroimaging efforts with clinical applications including tracking of neurodegenerative disease progression and potential use as a biomarker in the monitoring of a therapeutic response in large multi-site clinical trials.

Measurement of the hippocampus has been demonstrated to be reliable in prior work (Iglesias et al., 2015, 2016; Iglesias et al., 2013; Mueller et al., 2018; Whelan et al., 2016; Wisse et al., 2016; Worker et al., 2018; Zou et al., 2004), however, the subregional structures are challenging to accurately segment and measure in part due to their small size and lack of appropriate signal contrast with typical structural magnetic resonance imaging (MRI) acquisitions. Recent developments in acquisition hardware and sequence technology have increased various aspects of scan quality including resolution and signal-to-noise ratio (SNR) than previously possible (Whelan et al., 2016; Worker et al., 2018; Zeineh et al., 2001), overcoming some of these prior limitations. Furthermore, novel techniques have recently been developed for segmentation of the hippocampal subfields from MRI (Adler et al., 2014, 2018; Berron et al., 2017; DeKraker et al., 2019b; DeKraker et al., 2017; Giuliano et al., 2017; Iglesias et al., 2015; Pipitone et al., 2014; Sankar et al., 2017; Whelan et al., 2016; Winterburn et al., 2013; Wisse et al., 2016; Yushkevich et al., 2010a,b). We examine here the probabilistic atlas-based procedure released in the FreeSurfer processing stream (version 6.0; http://surfer.nmr.mgh.harvard.edu). This method has recently been demonstrated to provide robust discrimination between individuals with AD and cognitively healthy control participants (88% accuracy; Iglesias et al., 2015). These impressive results highlight the sensitivity and clinical utility of such procedures yet do not provide information about the reliability, a necessary parameter for measurement of longitudinal change as well as potential assessment of a treatment effect in a clinical trial.

Here we examined test-retest reliability at two time points within a Siemens Trio scanner and an additional time point acquired on a Siemens Prisma^{fit} (Trio upgrade) to determine the utility of hippocampal subfield measurement for longitudinal studies. The reliability of measures derived from automated morphometric procedures can be influenced by several sources of variance, including subject and instrument-related factors, such as field strength and scanner manufacturer. Longitudinal and multi-site studies face additional challenges associated with both subject and instrument-related factors, such as scanner upgrades and differences in software and hardware components (De Guio et al., 2016; Han et al., 2006; Jovicich et al., 2006). This work follows prior studies examining the reliability of FreeSurfer's automated hippocampal segmentation algorithm (Mueller et al., 2018; Tamnes et al., 2018; Whelan et al., 2016; Worker et al., 2018) with an assessment of measurements across scanner upgrade which has not previously been explored. The standard cross-sectional processing pipeline in FreeSurfer 6.0 was used for this work. Additionally, the recently released longitudinal analysis tool in FreeSurfer 6.0 was used, which references a within-subject template to enforce consistent segmentation results across time points and reduce the confounding effects associated with longitudinal analysis. The results demonstrate that the FreeSurfer longitudinal stream provides a more reliable measurement of hippocampal subfields than the standard processing, and supports the potential use of neuroimaging biomarkers to track disease progression and as outcome measures in clinical trials to test therapeutic response in conditions promoting hippocampal neurodegeneration.

2. Materials and methods

2.1. Participants

Eleven generally healthy young to middle-aged individuals recruited from the research community participated in this study (age between 22 and 55 years; mean: 30.2 years; SD: 9.43 years. 6 males, 5 females). All participants received a baseline and a follow-up scan within 2 months with a range of 7–50 days on a Siemens Trio scanner. Participants were then scanned approximately 2 months (with a range of 50–70 days) after their second Trio scan on the upgraded Siemens Prisma^{fit} scanner.

The Boston VA Medical Center Institutional Review Board approved this study and participants provided written informed consent.

2.2. Image acquisition

Each subject underwent 3 scan sessions, twice before and once after MRI scanner upgrade at approximately 2-month intervals, a total time span of about 4 months. The upgrade was from a Siemens 3 T Magnetom Trio scanner to a Magnetom Prisma^{fit}, which included the following major changes: main magnet (both are 3 T, Trio's length is 215 cm, Prisma^{fit} is 198 cm), gradient system (Trio 40 mT/m at 200 T/m/s, Prisma^{fit} 80 mT/m at 200 T/m/s), and Syngo software upgrade (Trio B17, Prisma^{fit} D13D; Siemens Medical Solutions, Erlangen, Germany). Although scanner upgrades are coordinated to make minimal changes so as not to invalidate data, scanner upgrade remains a significant image acquisition variable. The Siemens Trio and Prisma^{fit} are two widely used scanners in neuroimaging studies. Transitioning between the two is a major hardware and software upgrade where, essentially, the only thing that does not change is the main static magnet. Furthermore, sequences were not identical, which contributes additional variables affecting image acquisition. For each scan session, the acquisition included two high-resolution T1-weighted images using Magnetization-Prepared Rapid Gradient Echo (MPRAGE) volumes with a 20-channel phased-array head coil and approximately matched parameters (200 Hz/pixel bandwidth, flip angle = 7 deg, Trio: TR/TE/TI = 2.53 s/3.32 ms/1.1 s, Prisma^{fit}: TR/TE/TI = 2.53 s/3.35 ms/1.1 s). All scans were 3D sagittal acquisitions with 176 contiguous slices (imaging matrix = 256 × 176, in-plane resolution = 1 mm, slice thickness = 1 mm). Acquisition time for both sequences was 6:02 min.

2.3. Image analysis

2.3.1. Standard FreeSurfer processing pipeline

All T1-weighted images were visually inspected for motion artifact

and gray-white contrast. The single acquisition with less motion artifact from each scan session was used so as not to introduce noise from the additional volume. Cortical reconstruction and volumetric segmentation was performed with version 6.0 of the FreeSurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu). The technical details of these procedures are described in prior publications (Dale et al., 1999; Dale and Sereno, 1993; Fischl and Dale, 2000; Fischl et al., 2001; Fischl et al., 1999; Fischl et al., 2002, 2004; Han et al., 2006; Jovicich et al., 2006; Reuter et al., 2010; Reuter et al., 2012; Ségonne et al., 2004). Briefly, this processing includes motion correction and averaging (Reuter et al., 2010) of multiple volumetric T1 weighted images (when more than one is available), removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Ségonne et al., 2004), automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles) (Fischl et al., 2002; 2004), intensity normalization (Sled et al., 1998), tessellation of the gray matter white matter boundary, automated topology correction (Fischl et al., 2001; Ségonne et al., 2007), and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class (Dale et al., 1999; Dale and Sereno, 1993; Fischl and Dale, 2000). FreeSurfer morphometric procedures have been demonstrated to show good test-retest reliability across scanner manufacturers and across field strengths (Han et al., 2006; Reuter et al., 2012). After all subjects were run through the standard processing stream, the data was manually inspected and edited for accuracy of the gray/white and gray/pial surfaces. Minimal edits were necessary for these subjects. Images were then run through second reconstruction, beginning at the point where edits were applied.

2.3.2. Longitudinal processing

A significant challenge in longitudinal studies is the within-subject variability and lower reproducibility of repeated MRI scanning often due to subject or instrument-related factors (De Guio et al., 2016; Kruggel et al., 2010). However, using a longitudinal-specific approach can limit the variability and avoid the confounding effect associated with common methods, such as registering all follow-ups to the baseline scan. To extract reliable volume and thickness estimates, images were automatically processed with the longitudinal stream (Reuter et al., 2012) in FreeSurfer v6.0. Specifically, an unbiased within-subject template space and image is created using robust, inverse consistent registration (Reuter et al., 2010). Several processing steps, such as skull stripping, Talairach transforms, atlas registration as well as spherical surface maps and parcellations are then initialized with common information from the within-subject template, significantly increasing reliability and statistical power (Reuter et al., 2012). Using this method, a within-subject template is referenced to enforce consistent segmentation results across time points and thereby reducing the confounding effects associated with longitudinal analysis which improves the robustness and sensitivity of the overall analysis (Iglesias et al., 2016). The longitudinal processing was performed with the inclusion of scans from all three time points, and repeated for two time points both within scanner (Trio-A and Trio-B) and across scanners (Trio-B and Prisma^{fit}).

2.3.3. Hippocampal subfields

A pipeline for automated hippocampal subfield segmentation released as part of FreeSurfer v6.0 was applied to the reconstructed images produced by the cross-sectional pipeline, which yielded volumetric estimations of each subregion. FreeSurfer v6.0 has also released a hippocampal subfield segmentation algorithm for longitudinal segmentation that is applied to each within-subject template produced by the longitudinal pipeline wherein the processing steps are initialized for each time point with common information from the subject template. The details of these steps are described in the original paper of this method (Iglesias et al., 2016). Briefly, the algorithm for segmentation of individual subregions uses Bayesian inference based on observed image intensities and a probabilistic atlas built from a library of *in vivo* manual segmentations and ultra-high resolution (~0.1 mm isotropic) *ex vivo* labeled MRI data (Iglesias et al., 2016; Van Leemput, 2009). The longitudinal pipeline uses a binary mask of the hippocampus that has been extracted from the automated segmentation of each subject's base template using soft segmentation of the hippocampus. The resulting mesh is then deformed to the same within-subject automated segmentation. This deformed mesh is used to initialize the positions of nodes within the base template as well as the time points. Additionally, the whole brain segmentation is used to improve the estimation of Gaussian parameters for particular tissue classes. See Fig. 1 for visualization of hippocampal subfield segmentation and Fig. 2 for a 3D rendering.

A whole hippocampal volume estimate was also used from this tool based on the binary mask from the standard pipeline and a soft segmentation of this subfield pipeline. This volume estimate is different from the FreeSurfer standard whole brain segmentation, as the estimates were found to be more accurate for AD discrimination in the original paper (Iglesias et al., 2016).

2.4. Statistical methods

2.4.1. Percent volume difference and dice overlap

Mean percent volume differences and Dice overlap were calculated to determine volumetric correspondence of each subregion given by Equations (1) and (2).

$$\text{Percent volume difference} = \frac{|\text{A} - \text{B}|}{\left(\frac{\text{A}+\text{B}}{2}\right)} \times 100 \quad (1)$$

$$\text{Dice overlap} = \frac{|A| \cap |B|}{\left(\frac{A+B}{2}\right)} \times 100 \quad (2)$$

Note. Additional variable C was used to denote the third scan (A-B; B–C; A-C).

In these equations, A represents a given subfield measurement from the first scan on the Trio scanner, and B represents the same subfield from the second scan on the Trio. An additional variable C was used to represent the subfield measurement provided by the Prisma^{fit} scanner. Equation (1) was used to estimate the mean percent volume difference between three time points across the two scanners (Trio-A to Trio-B; Trio-B to Prisma^{fit}-C; Trio-A to Prisma^{fit}-C) where an optimal value of zero would indicate no difference between volumes and increasing values indicate greater volume difference.

Equation (2) was used to estimate the Dice overlap between the two scans, where an optimal value of 100 is achieved for identical volumes, and a decrease in values indicates less overlap. Dice overlap is a well-used metric for verifying some volumetric correspondence exists between the ground truth and the estimated labels. This was repeated for the two time points provided by the Trio scanner, and then again across the two scanners.

All subfields will be correlated with total hippocampal volume to some degree as these are dependent measures (Elman et al., 2019; Greenspan et al., 2016; Patel et al., 2017). However, we examined the correlation between volumes of left and right hippocampal subfields with ipsilateral hippocampal volume by calculating Pearson's correlation and p-values.

Additionally, we have included Bland-Altman plots to visualize the reliability of the subfields across time points and processing pipelines. These plots can be found in the supplementary material.

2.4.2. Intra-class correlation coefficient (ICC)

Percent volume difference and Dice overlap; however, are insufficient

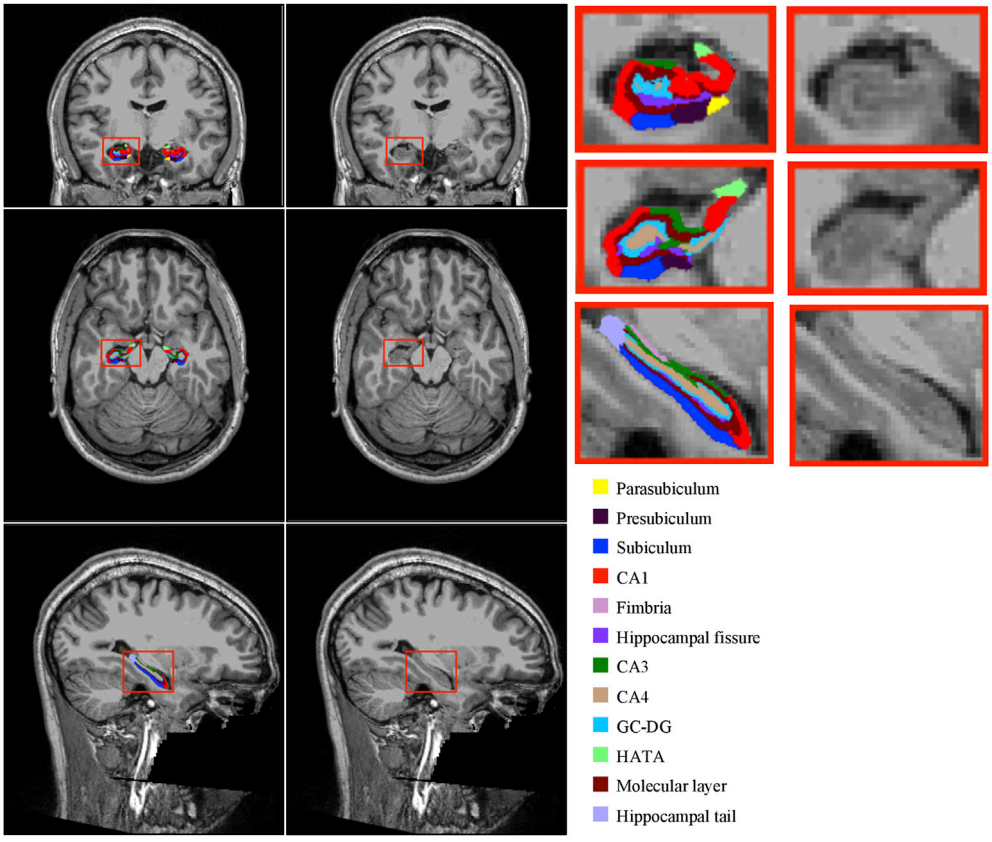


Fig. 1. *Visualization of the hippocampal subfield segmentations for a single subject and color key*. GC-DG (Granule cell layer of the dentate gyrus), HATA (Hippocampal-amygdaloid transitional area), CA (Cornu ammonis).

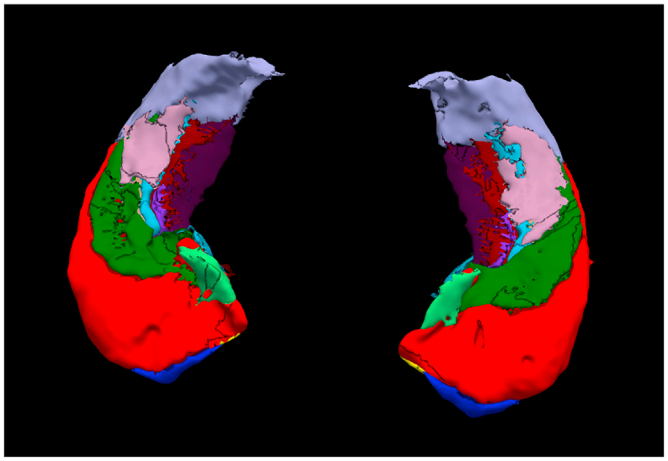


Fig. 2. 3D rendering of the hippocampal subfield segmentations for a single subject.

methods to examine reliability alone, because volume difference does not measure variability both within-subject and between subject, and Dice does not account for variability in ground truth labeling, which is essential to determine whether a method can be considered "good enough." In test-retest data, the intra-class coefficient (ICC) can be used to measure within-subject variability relative to between-subject variability. The third form of the ICC ($ICC_{3,1}$), as defined in previous literature (Shrout and Fleiss, 1979), was applied to each subfield by hemisphere to estimate the agreement of measures between the three scans across two scanners. This calculation was repeated to estimate agreement for two scans provided by the same scanner. The ICC analysis was modelled by a two-way mixed-effects model; random subject effects and fixed sessions effects, with absolute agreement. A statistical package "irr" designed for ICC analysis implemented using R was used to calculate ICC_{max} values from the mean volumes of each subregion (Gamer et al., 2012). In order to assess the reliability of the hippocampal segmentation pipeline both cross-sectionally and longitudinally, we have reported measures for both processing pipelines.

As a speculative analysis to determine the power of this segmentation tool in detecting effects on hippocampal atrophy, we evaluated power denoted as sample sizes required for a proposed therapeutic intervention to reduce rates of hippocampal atrophy between an Alzheimer and control group.

3. Results

3.1. Volumetric correspondence

Fig. 3 displays the percent volume difference processed with the standard FreeSurfer pipeline and the longitudinal pipeline including all three time points. Values represent the mean volume difference between each session comparison (Trio-A to Trio-B; Trio-B to $Prisma^{fit}$-C; Trio-A to $Prisma^{fit}$-C). The whole hippocampus, molecular layer, dentate gyrus, and CA1 show the most consistency across time points with less than 3% mean volume difference bilaterally between time points when processed through the longitudinal pipeline. While the fissure, parasubiculum, and fimbria show the least consistency with a greater than 5% mean volume difference between time points. The volume difference values from Trio-A to Trio-B are consistently lower than those across scanner, which suggests that within-scanner sessions performed better overall.

Fig. 4 displays the mean percent volume difference for two time points, both within-scanner and across-scanner processed through the standard and longitudinal FreeSurfer pipelines. The whole hippocampus,

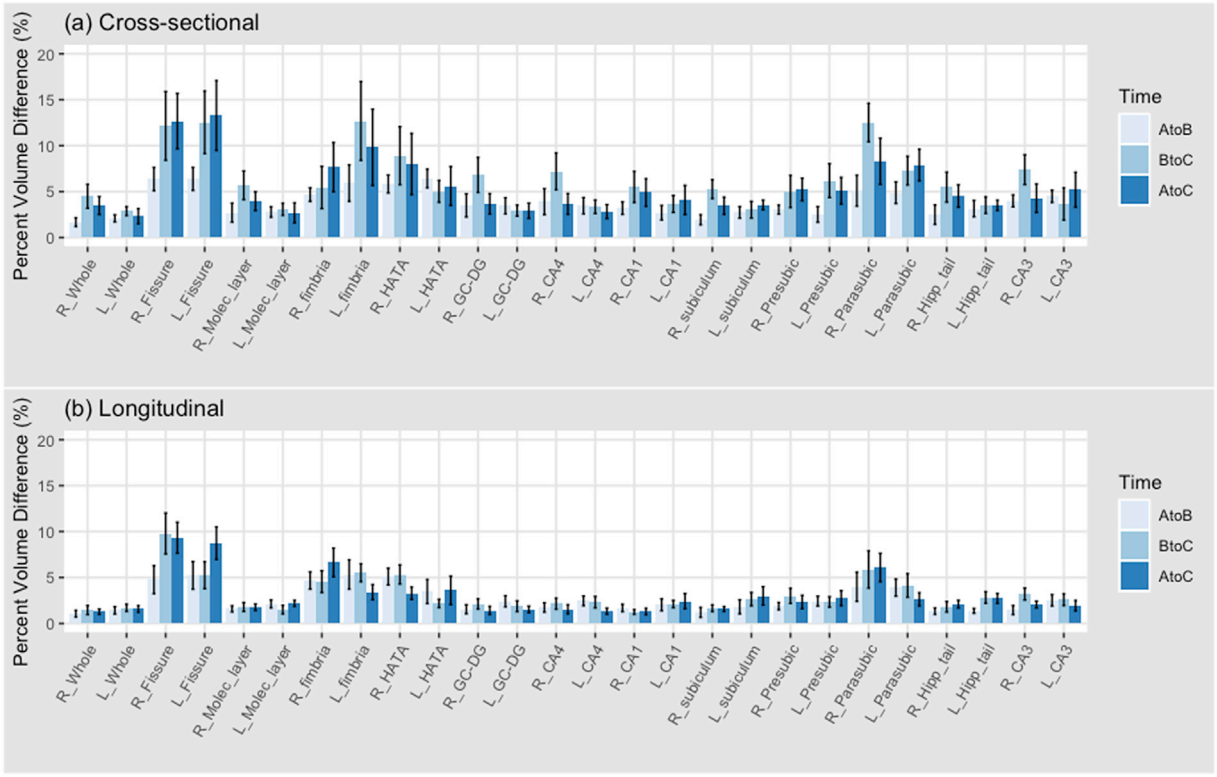


Fig. 3. *Longitudinal processing substantially increases test-retest reliability in all regions.* Mean percent volume difference and standard error bars for each subregion across three time points. (a) Cross-sectional (b) Longitudinal, healthy control subjects scanned at baseline (A) and 2 months (B) on a Siemens Magnetom Trio scanner, and 4 months (C) on a Siemens Magnetom Prisma^{fit} scanner. An optimal value of zero indicates no difference, therefore higher bars indicate worse performance. Trio-A to Trio-B bars (light blue) are consistently lower than darker blue bars, indicating within scanner performance is better than across scanner.

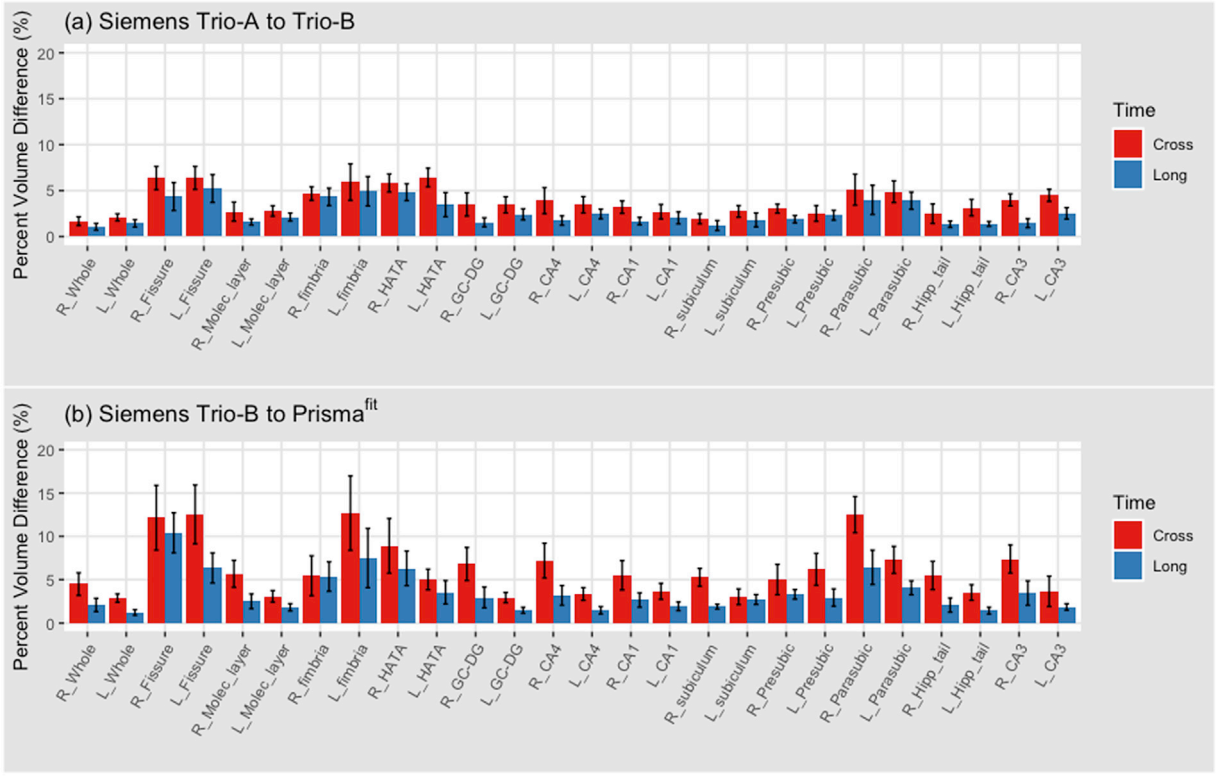


Fig. 4. *Longitudinal processing substantially increases test-retest reliability in all regions.* Mean percent volume difference and standard error bars for two time points. (Cross)ectional (red) vs (Long)itudinal (blue), healthy control subjects scanned at baseline and 2 months on a Siemens Magnetom Trio (a), and scanned at 2 months and 4 months from a Siemens Magnetom Trio to Siemens Magnetom Prisma^{fit} (b). An optimal value of zero indicates no difference, therefore higher bars indicate worse performance. Longitudinal processing (blue) produced lower volume difference values. The within scanner bars (a) are consistently lower than the across scanner volumes, indicating within scanner performance is better than across scanner. The inclusion of an additional scan may add noise and therefore the longitudinal estimates may be affected.

molecular layer, and dentate gyrus remained stable with the inclusion of two time points. However, the volume differences within scanner are consistently lower than those across scanners, indicating within scanner measurement performed better. Differences in percent volume difference across time points and processing between the left and right hippocampal subfields were not significant ($p > 0.05$). This suggests the inclusion of an additional scan may add noise and therefore the longitudinal estimates may be affected.

Fig. 5 displays the Dice overlap coefficients processed with the standard and longitudinal pipeline including all three time points. Values represent the Dice overlap between each session comparison. The whole hippocampus, subiculum, presubiculum, and hippocampal tail demonstrate the most consistency across time points with a greater than 75% overlap bilaterally between time points when processed through the cross-sectional and longitudinal pipeline. Only the fissure remains the least consistent with a less than 70% mean volume overlap between time points. The Dice coefficients consistently increased with longitudinal processing with most subregions achieving scores greater than 80%.

Fig. 6 displays the Dice overlap coefficients for two time points, both within-scanner and across-scanner processed through the standard and longitudinal FreeSurfer pipelines. The whole hippocampus, CA4, and hippocampal tail remained stable with the inclusion of two time points. The overlap consistently increased with longitudinal processing with most subregions achieving scores of 75% or greater. Differences in Dice overlap values across time points and processing between the left and right hippocampal subfields were not significant ($p > 0.05$). However, the Dice overlaps within scanner are consistently higher than those across scanners, indicating within scanner measurement performed better. This further suggests the inclusion of an additional scan may add noise and affect the longitudinal assessment.

Fig. 7 displays whole hippocampal plots for two time points within and across scanner using both processing pipelines. Whole hippocampal volumes were paired by hemisphere to observe the raw value reliability across scanner. The stronger correlation in the graphs plotting longitudinal processing shows significant stabilization of noise through this pipeline.

Table 1 reports correlation coefficients for regional associations of subfield volumes to ipsilateral hippocampal volume across processing pipelines with the inclusion of all three time points across scanners. Both the left and right subfield volumes were positively correlated with whole hippocampal volume. Bilaterally, the subiculum, CA1, and molecular layer demonstrate the strongest correlation with their respective hemispheric hippocampal volume ($r > 0.90$, $p < 0.001$), whereas the lowest correlation coefficients were the left and right fissure, and the right parasubiculum ($r < 0.50$, $p < 0.001$). Given that there is a broad range of correlation values (0.32–0.99), we observed a range of ways the subregions relate to the total hippocampal volume.

3.2. Test-retest reliability

Reliability was generally high among subregions ranging from 0.74 to 0.98 cross-sectionally and 0.92–0.99 after longitudinal processing with the inclusion of all three time points (Table 2). Variability of percent volume difference was observed in regions while maintaining high reliability in the molecular layer, dentate gyrus, and whole hippocampus (>0.98), and moderately high in the fissure, parasubiculum, fimbria, and CA3 (>0.92). The variability observed is assumed to be due to sensitivity to inter-subject variation. With the inclusion of all three time points, all regions achieved ICC scores greater than 0.90 with longitudinal processing. ICC values of the longitudinal stream were significantly higher than for the cross-sectional stream in the right CA1, fissure, right dentate gyrus, right CA3, right CA4, fimbria, right HATA and whole hippocampus. We observed higher reliability in the left hippocampal subregions compared to the right, which was statistically significant different for

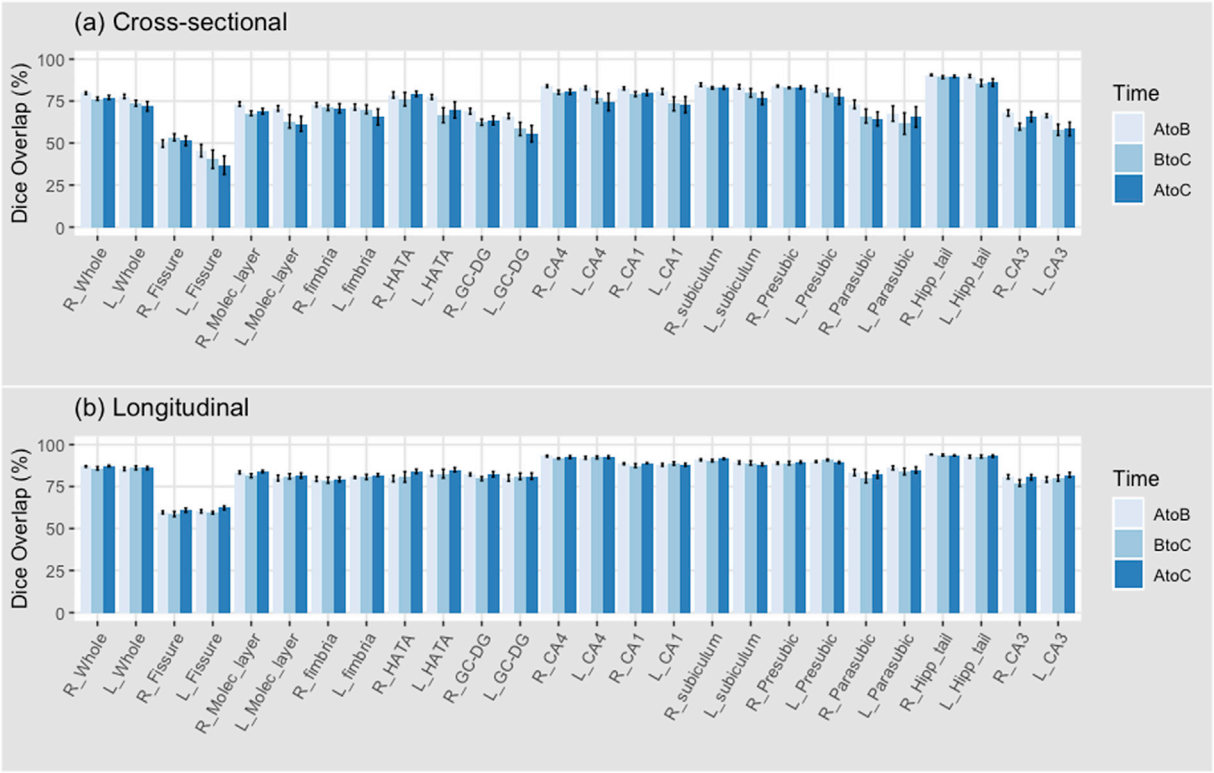


Fig. 5. *Longitudinal processing substantially increases test-retest reliability in all regions.* Dice overlap and standard error bars for each subregion across three time points. (a) Cross-sectional (b) Longitudinal, healthy control subjects scanned at baseline (A) and 2 months (B) on a Siemens Magnetom Trio scanner, and 4 months (C) on a Siemens Magnetom Prisma^{fit} scanner. An optimal value of 100 indicates no difference, therefore lower bars indicate worse performance.

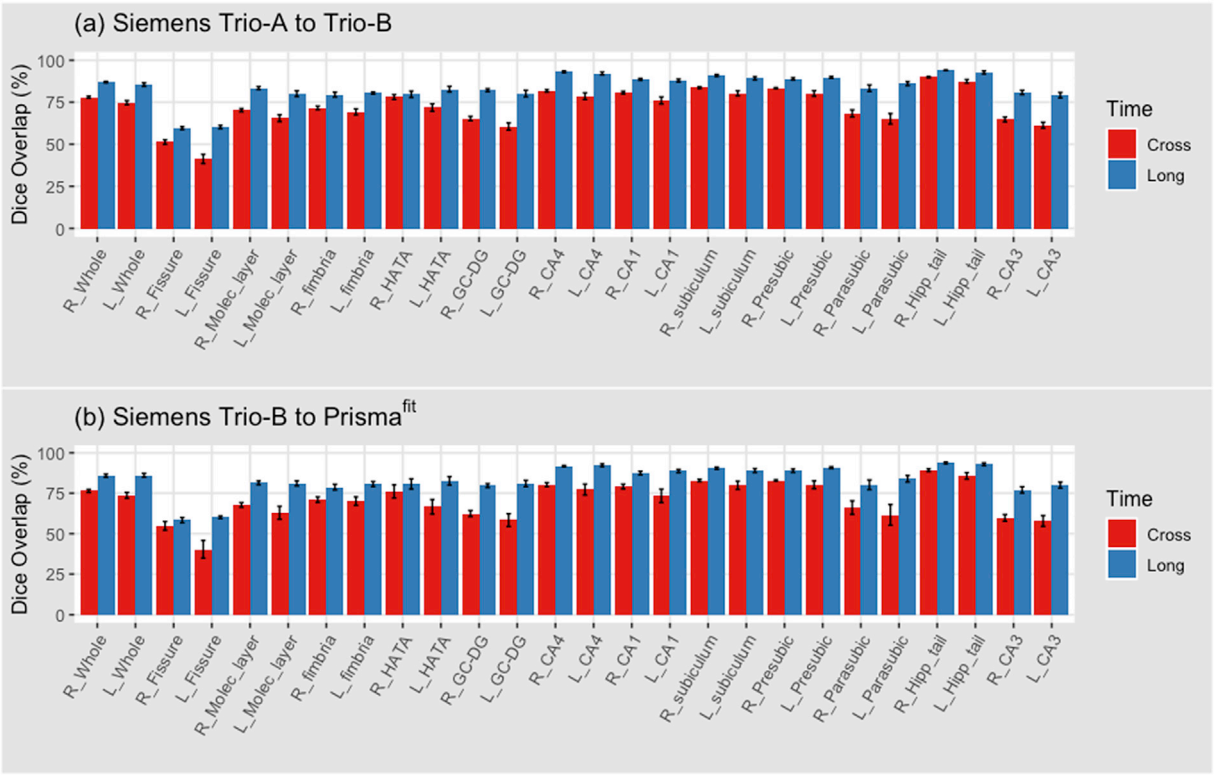


Fig. 6. *Longitudinal processing substantially increases test-retest reliability in all regions*. Dice overlap and standard error bars for two time points. (Cross)ectional (red) vs (Long)itudinal (blue), healthy control subjects scanned at baseline and 2 months on a Siemens Magnetom Trio (a), and scanned at 2 months and 4 months from a Siemens Magnetom Trio to Siemens Magnetom Prisma^{fit} (b). An optimal value of 100 indicates no difference, therefore lower bars indicate worse performance. The within scanner bars (a) are consistently higher than the across scanner volumes, indicating within scanner performance is better than across scanner. The inclusion of an additional scan may add noise and therefore the longitudinal estimates may be affected.

cross-sectional ($p < 0.05$). However, there was no significant difference in reliability coefficients between hemispheres for time points processed through the longitudinal pipeline ($p > 0.05$).

With the inclusion of only two time points across scanner (Table 3), subregions were relatively stable with most achieving ICC scores over 0.90, with the exception of the right fissure and right CA3 remaining below 0.80. Although these regions were less consistent, they also increased significantly with longitudinal processing. There was no statistically significant difference of reliability coefficients between hemispheres for two time points processed through the cross-sectional or longitudinal pipeline ($p > 0.05$).

To visualize the reliability of the subfield volume estimations across time points and processing pipelines, we have included Bland-Altman plots in the supplementary material. Each subregion is plotted by mean and volume difference between time points across three time points, processed with the cross-sectional and longitudinal pipelines. The longitudinal pipeline substantially reduced volume differences across all subregions. With volume differences and mean remaining in consistent locations across time points. However, the volume differences were often higher for assessments across scanner, which again suggests the inclusion of an additional scan may add noise.

3.3. Sample size estimation

To determine the power of a FreeSurfer segmentation-based atrophy assessment in detecting effects on hippocampal volume loss over time, we estimated the sample size needed in a proposed therapeutic intervention of 25% reduction in atrophy over a 12-month trial with 6-month assessment intervals at 80% confidence and 0.05 significance. This secondary analysis was performed for whole hippocampal volume only as there is a limited number of publications reporting individual subfield atrophy rates. A power calculation was performed using a mix-effects regression model for the outcome variable (assumed mean rate of decline) as a function of time. The estimated sample sizes are presented in Table 4. When processed using the longitudinal method, we observed significantly reduced percent volume difference and standard error between time points compared to the cross-sectional method. Furthermore, the required sample size for a hypothetical therapeutic intervention also reflected this difference. Given a variation of 3% difference for the whole hippocampus bilaterally when using the longitudinal method while considering standard error, the FreeSurfer segmentation method should be able to detect about a 1.5% difference between diagnostic groups with 80% power at a 0.05 level of significance.

4. Discussion

To our knowledge, this is the first study to assess the test-retest reliability using percent volume difference, Dice overlap, and ICC of automated hippocampal subfield segmentation applied to cross-sectional and longitudinal data using FreeSurfer's pipeline with scanner upgrade consisting of three separate time points across two scanners. Most of the hippocampal subregions were found to be highly stable and all regions achieve high ICC values after longitudinal processing and scanner upgrade (ICC>0.9). These results are consistent with previous findings assessing reliability of FreeSurfer's automated hippocampal subfield segmentation pipeline (Mueller et al., 2018; Tamnes et al., 2018; Whelan et al., 2016; Worker et al., 2018), with substantially increased reliability due to the updated processing in the most stable regions both within scanner and across scanners, whilst the fissure and fimbria were the least stable. The hippocampal fissure lies between the molecular layer and the dentate gyrus, an area difficult to manually segment even in histological studies as it is also vulnerable to signal loss and distortion especially in

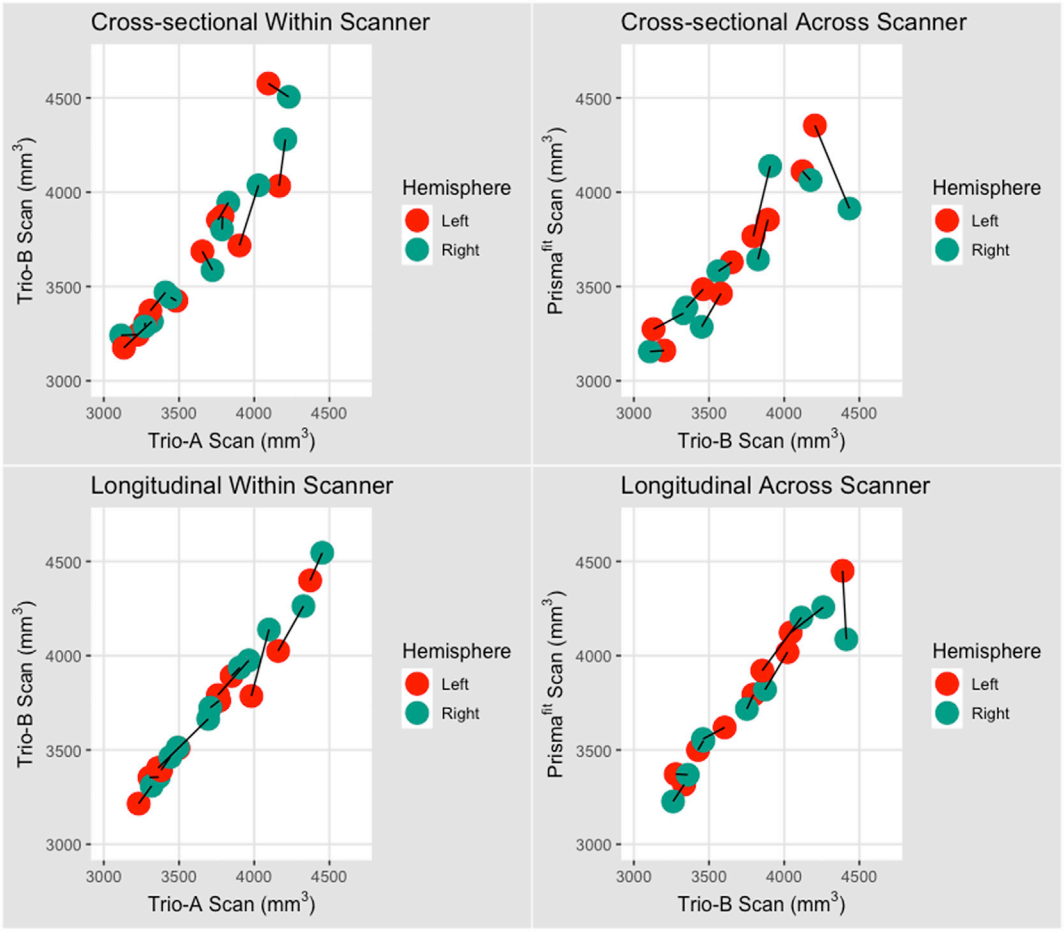


Fig. 7. *Longitudinal processing substantially decreases noise in whole hippocampal volume across scanner.* Plots of correlation between whole hippocampal volumes by hemisphere paired for each subject from a Siemens Trio scanner to a Siemens Prisma^{fit} scanner from FreeSurfer segmentation. Left hemisphere is presented in red and the right hemisphere in green.

Table 1
Correlation coefficients for hippocampal subfield volume associations with whole hippocampal volume at three time points across scanner upgrade (Trio-A to Trio-B to Prisma^{fit}) processed cross-sectionally and longitudinally. All subfield volumes showed a significant correlation between ipsilateral hippocampal volume ($p < 0.001$).

Subregion	Cross-sectional		Longitudinal	
	Left R^2	Right R^2	Left R^2	Right R^2
Hippocampal tail	0.75	0.77	0.86	0.72
Subiculum	0.93	0.94	0.96	0.96
CA1	0.89	0.93	0.96	0.91
Fissure	0.21	0.52	0.10	0.32
Presubiculum	0.82	0.90	0.70	0.89
Parasubiculum	0.79	0.58	0.75	0.45
Molecular layer	0.97	0.98	0.96	0.99
GC-DG	0.80	0.89	0.58	0.70
CA3	0.67	0.91	0.65	0.86
CA4	0.83	0.87	0.57	0.70
Fimbria	0.39	0.57	0.75	0.68
HATA	0.60	0.58	0.86	0.63

lower contrast images which contributes to the lower reliability (Olsen et al., 2019; Van Leemput, 2009; Whelan et al., 2016; Worker et al., 2018; Yushkevich et al., 2015; Yushkevich et al., 2010a,b). Other regions with observed lower reliability are among the smallest of the subregions, making them susceptible to partial volume effects.

Although standard 1 mm resolution is a likely common nature of acquisition and is widely supported in open source software suites for image processing (e.g., FreeSurfer), these findings suggest that there are additional factors to consider, primarily how this resolution can be accurately localized from *ex vivo* data. The segmentation version used in this study has improved by considering differing contrast properties in *ex vivo* and *in vivo* data, although questions remain as to the appropriate protocols for segmentation of the hippocampus, even at the cytoarchitectonic level (DeKraker et al., 2019b; DeKraker et al., 2019a; de Flores et al., 2019; Olsen et al., 2019; Yushkevich et al., 2015). The progress to date in consensus for histological subfield determination demonstrates the challenges of applying such labeling at the *in vivo* scale. Without an established method to map *in vivo* to *ex vivo* data precisely due to differing contrast properties of tissue between resolution, or a gold standard method obtained by non-imaging methods such as histology, validation must be an assessment of reliability or reproducibility (Zou et al., 2004). Some of these factors for further validation can be addressed by including an additional T2-weighted or proton-density volume co-registered with the standard resolution data to ensure accurate labeling and possibly improve the degree of overlap in the subregions (Iglesias et al., 2015). The methods used for test-retest reliability, however, do not account for variability in the ground truth labeling, which is essential to determine whether a method can be considered "good enough." Power assessment of the current volume estimation can be achieved by the inclusion of an additional dataset, or assessing utility of the subfields as seed and target regions on diffusion and functional MRI (Iglesias et al., 2016).

This data is preliminary and will require further validation with a larger dataset that includes data across different sites and resolution,

Table 2
Intraclass correlation coefficients for hippocampal subregion volumes at three time points across scanner upgrade (Trio-A to Trio-B to Prisma^{fit}) processed cross-sectionally and longitudinally with 95 confidence intervals.

Subregion	Hemi	Cross-sectional			Longitudinal		
		Mean ICC	Lower ICC	Upper ICC	Mean ICC	Lower ICC	Upper ICC
Hippocampal tail	Left	0.98	0.94	0.99	0.99	0.96	0.99
	Right	0.97	0.90	0.99	0.99	0.99	0.99
Subiculum	Left	0.97	0.91	0.99	0.98	0.92	0.99
	Right	0.96	0.87	0.99	0.99	0.97	0.99
CA1	Left	0.93	0.77	0.98	0.96	0.87	0.99
	Right	0.86	0.58	0.97	0.96	0.88	0.99
Fissure	Left	0.89	0.64	0.97	0.97	0.90	0.99
	Right	0.78	0.33	0.95	0.92	0.75	0.98
Presubiculum	Left	0.92	0.76	0.98	0.99	0.97	0.99
	Right	0.93	0.80	0.98	0.98	0.95	0.99
Parasubiculum	Left	0.96	0.87	0.99	0.99	0.96	0.99
	Right	0.91	0.65	0.98	0.96	0.87	0.99
Molecular layer	Left	0.96	0.87	0.99	0.98	0.95	0.99
	Right	0.91	0.72	0.98	0.98	0.94	0.99
GC-DG	Left	0.97	0.92	0.99	0.99	0.97	0.99
	Right	0.88	0.63	0.97	0.99	0.96	0.99
CA3	Left	0.91	0.72	0.98	0.99	0.96	0.99
	Right	0.74	0.17	0.94	0.98	0.93	0.99
CA4	Left	0.97	0.90	0.99	0.99	0.97	0.99
	Right	0.84	0.51	0.96	0.99	0.96	0.99
Fimbria	Left	0.86	0.56	0.97	0.97	0.89	0.99
	Right	0.93	0.79	0.98	0.96	0.86	0.99
HATA	Left	0.94	0.80	0.99	0.95	0.84	0.99
	Right	0.86	0.57	0.96	0.95	0.86	0.99
Whole	Left	0.96	0.89	0.99	0.99	0.95	0.99
	Right	0.92	0.76	0.98	0.99	0.96	0.99

Note. GC-DG (Granule cell layer of the dentate gyrus), HATA (Hippocampal-amygdaloid transitional area). Whole hippocampus represented the measure of whole hippocampal volume produced by the pipeline.

Table 3
Intraclass correlation coefficients for hippocampal subregion volumes at two time points across scanner upgrade (Trio-B to Prisma^{fit}) processed cross-sectionally and longitudinally with 95 confidence intervals.

Subregion	Hemi	Cross-sectional			Longitudinal		
		Mean ICC	Lower ICC	Upper ICC	Mean ICC	Lower ICC	Upper ICC
Hippocampal tail	Left	0.96	0.82	0.99	0.99	0.97	0.99
	Right	0.91	0.60	0.98	0.97	0.88	0.99
Subiculum	Left	0.94	0.72	0.99	0.97	0.87	0.99
	Right	0.87	0.48	0.97	0.99	0.94	0.99
CA1	Left	0.91	0.61	0.98	0.98	0.92	0.99
	Right	0.67	0.21	0.92	0.94	0.75	0.99
Fissure	Left	0.71	0.36	0.94	0.93	0.67	0.98
	Right	0.43	0.15	0.87	0.78	0.35	0.95
Presubiculum	Left	0.82	0.21	0.96	0.97	0.85	0.99
	Right	0.87	0.45	0.97	0.96	0.81	0.99
Parasubiculum	Left	0.93	0.71	0.98	0.98	0.92	0.99
	Right	0.82	0.08	0.96	0.89	0.37	0.98
Molecular layer	Left	0.93	0.73	0.99	0.98	0.85	0.99
	Right	0.74	0.01	0.94	0.95	0.76	0.99
GC-DG	Left	0.97	0.85	0.99	0.99	0.94	0.99
	Right	0.62	0.37	0.91	0.85	0.41	0.97
CA3	Left	0.89	0.54	0.98	0.99	0.88	0.99
	Right	0.32	0.22	0.85	0.54	0.25	0.90
CA4	Left	0.95	0.78	0.99	0.99	0.94	0.99
	Right	0.51	0.14	0.89	0.80	0.02	0.95
Fimbria	Left	0.65	0.48	0.92	0.85	0.29	0.97
	Right	0.89	0.51	0.98	0.88	0.48	0.97
HATA	Left	0.94	0.06	0.99	0.94	0.77	0.99
	Right	0.70	0.30	0.93	0.83	0.25	0.96
Whole	Left	0.96	0.89	0.99	0.99	0.95	0.99
	Right	0.92	0.76	0.98	0.99	0.96	0.99

which is currently underway in validation studies. However, it is important for studies that combine data acquired across multiple sites to understand and adjust for instrument-related differences, such as software and hardware components, as well as scanner manufacturer and field strength (Han et al., 2006; Jovicich et al., 2006). These results suggest there are additional factors to consider that may influence reliability. Significant differences between acquisitions within-scanner can be primarily assumed to be due to the algorithm, whereas significant differences between the within-scanner values (Trio to Trio) and the across scanner values (Trio to Prisma^{fit}) can be assumed to be due to scanner. Subsequent studies could model this bias and use it as a correction in analyses, or use a scanner covariate in a model. To limit the

Table 4
Estimations of sample size for each processing method in a proposed therapeutic intervention trial for Alzheimer's disease to detect a 25% reduction in atrophy with 80% power and $\alpha = 0.05$. Based on MRI scans at baseline, 6 months, and 12 months. Processed cross-sectionally and longitudinally.

Study Design	Cross-sectional	Longitudinal
Two scans; 0–6 months	1 440	505
Two scans; 0–12 months	467	251
Three scans; 0-6-12 months	470	242

sources of variance introduced by acquisition on multiple scanner platforms, having matched groups on both scanners can reduce the bias introduced by group and scanner. Having large sample sizes on both scanners can also limit the variability and analyses can be statistically modelled correctly. Additionally, the lack of sub-mm resolution in standard imaging procedures is a significant limitation and regions with lower reliability should be used cautiously. Previous publications discuss such procedures with regard to reliability of cortical thickness and subcortical volume measures (Han et al., 2006; Jovicich et al., 2006); however, it would be beneficial for both study design and interpretation to apply these methods to subfields of smaller limbic regions that are more susceptible to partial volume effects that may be influenced by the factors outlined above.

Despite these limitations, Iglesias and colleagues have improved upon an already robust atlas, which continues to provide additional information regarding individual subregions. Although most regions were stable and remained reliable after longitudinal processing and scanner upgrade, some regions varied in their reliability.

Application of this FreeSurfer software suite tool has demonstrated increased sensitivity and reliability of classification between healthy and neurodegenerative disorders such as AD than using a whole hippocampal volume assessment (Iglesias et al., 2016; Mueller et al., 2018; Worker et al., 2018). There is increasing interest in volumetric studies reporting higher rates of hippocampal volume loss in patients with Alzheimer's disease (AD) than in elderly controls (Ledig et al., 2018; Schuff et al., 2009; van der Flier et al., 2004; Zhao et al., 2019). However, global hippocampal volumetry has demonstrated moderate sensitivity and low specificity to AD diagnostic classification. Therefore, measurement of the substructures has become of great interest to clinical aging studies of therapeutic interventions to not only increase sensitivity to follow the progression of atrophy, but to also evaluate the atrophy in different substructures (Bocchetta et al., 2018; La Joie et al., 2013; Ledig et al., 2018; Maruszak and Thuret, 2014; Wolz et al., 2010; Yushkevich et al., 2010a,b; Zhao et al., 2019). In order to establish the clinical utility of a FreeSurfer subfield segmentation-based atrophy estimation as well as index the efficacy of a therapeutic intervention, it is necessary to consider effect size when interpreting volumetric assessments. When comparing controls and a mild cognitive impairment (MCI) or AD population, previous studies have determined the mean rate of atrophy of the whole hippocampus on the standard outcome for AD treatment trials to be around 4% per year (Ard and Edland, 2011; Caroli et al., 2015; de Flores et al., 2015; Holland et al., 2009; Ledig et al., 2018; Schuff et al., 2009; Wolz et al., 2010; Yushkevich et al., 2010a,b). Although our results are consistent with previous work showing that power of a hippocampal volume assessment increases with greater inter-scan interval (Caroli et al., 2015; Schuff et al., 2009), this is caveated by the fact that these results are based on data from a small sample size of generally young, healthy adults, and do not report power for each subregion. Future work would require subregional measurements in a large cohort of older adults of MCI or AD populations, preferably longitudinal in nature, to validate clinical utility. Results from such work would establish sensitivity and power within each region with a population that is particularly vulnerable. Furthermore, future directions would include assessing inter-scan interval effect sizes which can create negligent power increases, and the inclusion of other biomarkers such as ApoE4 status which can directly influence statistical power (Ard and Edland, 2011; Caroli et al., 2015; Schuff et al., 2009). Our results suggest that assessment of hippocampal subfield volume is attainable and given the limited variability in measures across scanner upgrade, this could apply to multicenter studies. Previous literature has demonstrated distinct functional properties and vulnerabilities of the substructures, including the dentate gyrus where neurogenesis has been shown to persist into adulthood, and therapies such as exercise and drug intervention have shown to increase neurogenesis (Leuner and Gould, 2010; Mcewen, 1999). Automated segmentation procedures offer the opportunity to reliably measure neurogenesis and regional susceptibility to therapeutic interventions.

Measuring the substructures that make up the hippocampus has been a key challenge in neuroimaging research due to their small size, signal loss in the medial temporal regions of the brain, and low contrast in sequences, therefore studies have been limited to modeling the hippocampus as a homogenous structure (Schuff et al., 2009; Van Leemput, 2009; Wisse et al., 2016; Worker et al., 2018; Yushkevich et al., 2015; Zeineh et al., 2001), or manually segmenting the regions (Adler et al., 2014, 2018; Berron et al., 2017; Di Paola et al., 2008; Hsu et al., 2002; Wisse et al., 2016). Modeling the hippocampus as a homogenous structure can sacrifice critical information, while manual tracing is often labor-intensive and inconsistent across studies. Therefore, a reliable automated segmentation procedure would have valuable applications to disease progression and clinical trials designed to assess the effects of pharmacological intervention.

5. Conclusion

The results presented here reflect the test-retest reliability of automated hippocampal subfield measures estimated from T1-weighted scans. Using intra-class correlation coefficients, percent volume difference, and Dice overlap, we were able to quantify the reliability of the volumes, and our results show most regions are highly stable with very small difference between subjects and sessions. However, we found that inclusion of additional scans influenced the reliability. Within scanner reliability (Trio-A to Trio-B) was worse when including the Prisma^{fit} scan in the longitudinal processing compared to only including the Trio scans in the processing. These results may suggest the need for matching of longitudinal points across subjects for longitudinal studies.

In conclusion, the results indicate that the methods applied are robust, and with further validation, could support the potential use in clinical trials to measure therapeutic response in conditions promoting hippocampal neurogenesis.

Funding

This research was supported by the Neuroimaging Research for Veterans Center (NeRVe) and the Translational Research Center for TBI and Stress Disorders (TRACTS), a VA Rehabilitation Research and Development Traumatic Brain Injury Center of Excellence (B3001-C). Juan E. Iglesias is supported by the European Research Council (ERC) (starting grant number 677697; project BUNGEE-TOOLS).

Author contributions

Emma M Brown: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper,

Meghan E. Pierce: Conceived and designed the analysis, Contributed data or analysis tools, Wrote the paper.

Dustin C. Clark: Collected the data, Contributed data or analysis tools.

Bruce R. Fischl: Conceived and designed the analysis, Contributed data or analysis tools, Wrote the paper.

Juan E. Iglesias: Conceived and designed the analysis, Contributed data or analysis tools, Wrote the paper.

William P. Milberg: Conceived and designed the analysis,

Contributed data or analysis tools, Wrote the paper.

Regina E. McGlinchey: Conceived and designed the analysis, Contributed data or analysis tools, Wrote the paper.

David H. Salat: Conceived and designed the analysis, Contributed data or analysis tools, Wrote the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116563.

References

Adler, D.H., Pluta, J., Kadivar, S., Craige, C., Gee, J.C., Avants, B.B., Yushkevich, P.A., 2014. Histology-derived volumetric annotation of the human hippocampal subfields in postmortem MRI. Neuroimage 84, 505–523. https://doi.org/10.1016/j.neuroimage.2013.08.067.

Adler, D.H., Wisse, L.E.M., Ittyerah, R., Pluta, J.B., Ding, S.-L., Xie, L., et al., 2018. Characterizing the human hippocampus in aging and Alzheimer's disease using a computational atlas derived from ex vivo MRI and histology. Proc. Natl. Acad. Sci. 115 (16), 4252–4257. https://doi.org/10.1073/pnas.1801093115.

Ard, M.C., Edland, S.D., 2011. Power calculations for clinical trials in Alzheimer's disease. J. Alzheimer's Dis.: JAD 26 (Suppl. 3), 369–377. https://doi.org/10.3233/JAD-2011-0062.

Benarroch, E.E., 2013. Adult neurogenesis in the dentate gyrus: general concepts and potential implications. Neurology 81 (16), 1443–1452. https://doi.org/10.1212/WNL.0b013e3182a9a156.

Berron, D., Vieweg, P., Hochkeppler, A., Pluta, J.B., Ding, S.-L., Maass, A., et al., 2017. A protocol for manual segmentation of medial temporal lobe subregions in 7Tesla MRI. Neuroimage: Clinical 15, 466–482. https://doi.org/10.1016/j.nicl.2017.05.022.

Bocchetta, M., Iglesias, J.E., Scelsi, M.A., Cash, D.M., Cardoso, M.J., Modat, M., et al., 2018. Hippocampal subfield volumetry: differential pattern of atrophy in different forms of genetic frontotemporal dementia. J. Alzheimer's Dis. 64 (2), 497–504. https://doi.org/10.3233/JAD-180195.

Brickman, A.M., Stern, Y., Small, S.A., 2011. Hippocampal subregions differentially associate with standardized memory tests. Hippocampus 21 (9), 923–928. https://doi.org/10.1016/j.bbr.2010.07.006.

Caroli, A., Prestia, A., Wade, S., Chen, K., Ayutyanont, N., Landau, S.M., et al., 2015. Alzheimer disease biomarkers as outcome measures for clinical trials in MCI. Alzheimers Dis. Assoc. Disord. 29 (2), 101–109. https://doi.org/10.1097/WAD.0000000000000071.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9 (2), 179–194. https://doi.org/10.1006/nimg.1998.0395.

Dale, A.M., Sereno, M.I., 1993. Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. J. Cogn. Neurosci. 5 (2), 162–176. https://doi.org/10.1162/jocn.1993.5.2.162.

de Flores, R., Berron, D., Ding, S.-L., Ittyerah, R., Pluta, J.B., Xie, L., et al., 2019. Characterization of hippocampal subfields using ex vivo MRI and histology data: lessons for in vivo segmentation. Hippocampus. https://doi.org/10.1002/hipo.23172, 31675165.

de Flores, R., La Joie, R., Chételat, G., 2015. Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease. Neuroscience 309, 29–50. https://doi.org/10.1016/j.neuroscience.2015.08.033.

De Guio, F., Jouvent, E., Biessels, G.J., Black, S.E., Brayne, C., Chen, C., et al., 2016. Reproducibility and variability of quantitative magnetic resonance imaging markers in cerebral small vessel disease. J. Cereb. Blood Flow Metab. 36 (8), 1319–1337. https://doi.org/10.1177/0271678X16647396.

DeKraker, J., Lau, J.C., Ferko, K.M., Khan, A.R., Köhler, S., 2019a. Hippocampal morphology and cytoarchitecture in the 3D BigBrain [Preprint]. https://doi.org/10.1101/599571.

DeKraker, J., Lau, J.C., Ferko, K.M., Khan, A.R., Köhler, S., 2019b. Hippocampal subfields revealed through unfolding and unsupervised clustering of laminar and morphological features in 3D BigBrain. Neuroimage 116328. https://doi.org/10.1016/j.neuroimage.2019.116328.

DeKraker, Jordan, Ferko, K.M., Lau, J.C., Köhler, S., Khan, A.R., 2017. Unfolding the hippocampus: An intrinsic coordinate system for subfield segmentations and quantitative mapping [Preprint]. https://doi.org/10.1101/146878.

Di Paola, M., Caltagirone, C., Fadda, L., Sabatini, U., Serra, L., Carlesimo, G.A., 2008. Hippocampal atrophy is the critical brain change in patients with hypoxic amnesia. Hippocampus 18 (7), 719–728. https://doi.org/10.1002/hipo.20432.

Elman, J.A., Panizzon, M.S., Gillespie, N.A., Hagler, D.J., Fennema-Notestine, C., Eyler, L.T., et al., 2019. Genetic architecture of hippocampal subfields on standard resolution MRI: how the parts relate to the whole. Hum. Brain Mapp. 40 (5), 1528–1540. https://doi.org/10.1002/hbm.24464.

Ergorul, C., Eichenbaum, H., 2004. The Hippocampus and memory for "What," "Where," and "When. Learn. Mem. 11 (4), 397–405. https://doi.org/10.1101/lm.73304.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. U. S. A. 97 (20), 11050–11055. https://doi.org/10.1073/pnas.200033797.

Fischl, B., Liu, A., Dale, A.M., 2001. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. IEEE Trans. Med. Imaging 20 (1), 70–80. https://doi.org/10.1109/42.906426.

Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M., 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. 8 (4), 272–284.

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

Fischl, Bruce, van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., et al., 2004. Automatically parcellating the human cerebral cortex. Cerebr. Cortex 14 (1), 11–22. New York, N.Y.: 1991.

Fortin, N.J., Agster, K.L., Eichenbaum, H.B., 2002. Critical role of the hippocampus in memory for sequences of events. Nat. Neurosci. 5 (5), 458–462. https://doi.org/10.1038/nn834.

Gamer, M., Lemon, J., Singh, I.F.P., 2012. Irr: various coefficients of interrater reliability and agreement (Version 0.84). Retrieved from. https://CRAN.R-project.org/package=irr.

Giuliano, A., Donatelli, G., Cosottini, M., Tosetti, M., Retico, A., Fantacci, M.E., 2017. Hippocampal subfields at ultra high field MRI: an overview of segmentation and measurement methods. Hippocampus 27 (5), 481. https://doi.org/10.1002/hipo.22717.

Greenspan, K.S., Arakelian, C.R., van Erp, T.G.M., 2016. Heritability of hippocampal formation sub-region volumes. J. Neurol. Neurosci. 7 (6) https://doi.org/10.21767/2171-6625.1000159.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., et al., 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://doi.org/10.1016/j.neuroimage.2006.02.051.

Hoge, J., Kesner, R., 2007. Role of CA3 and CA1 subregions of the dorsal hippocampus on temporal processing of objects. Neurobiol. Learn. Mem. 88 (2), 225–231. https://doi.org/10.1016/j.nlm.2007.04.013.

Holland, D., Brewer, J.B., Hagler, D.J., Fennema-Notestine, C., Dale, A.M., Alzheimer's Disease Neuroimaging Initiative, 2009. Subregional neuroanatomical change as a biomarker for Alzheimer's disease. Proc. Natl. Acad. Sci. U.S.A. 106 (49), 20954–20959. https://doi.org/10.1073/pnas.0906053106.

Horner, A.J., Doeller, C.F., 2017. Plasticity of hippocampal memories in humans. Curr. Opin. Neurobiol. 43, 102–109. https://doi.org/10.1016/j.conb.2017.02.004.

Horovitz, O., Richter-Levin, G., 2015. Dorsal periaqueductal gray simultaneously modulates ventral subiculum induced-plasticity in the basolateral amygdala and the nucleus accumbens. Front. Behav. Neurosci. 9 https://doi.org/10.3389/fnbeh.2015.00053.

Hsu, Y.-Y., Schuff, N., Du, A.-T., Mark, K., Zhu, X., Hardin, D., Weiner, M.W., 2002. Comparison of automated and manual MRI volumetry of Hippocampus in normal aging and dementia. J. Magn. Reson. Imaging: JMRI 16 (3), 305–310. https://doi.org/10.1002/jmri.10163.

Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., et al., 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.

Iglesias, J.E., Sabuncu, M.R., Leemput, K.V., 2013. Improved inference in Bayesian segmentation using Monte Carlo sampling: application to hippocampal subfield volumetry. Med. Image Anal. 17 (7), 766–778. https://doi.org/10.1016/j.media.2013.04.005.

Iglesias, J.E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., Reuter, M., Alzheimer's Disease Neuroimaging Initiative, 2016. Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. Neuroimage 141, 542–555. https://doi.org/10.1016/j.neuroimage.2016.07.020.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., et al., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 30 (2), 436–443. https://doi.org/10.1016/j.neuroimage.2005.09.046.

Kruggel, F., Turner, J., Muftuler, L.T., Alzheimer's Disease Neuroimaging Initiative, 2010. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. Neuroimage 49 (3), 2123–2133. https://doi.org/10.1016/j.neuroimage.2009.11.006.

La Joie, R., Perrotin, A., de La Sayette, V., Egret, S., Doeuvre, L., Belliard, S., et al., 2013. Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia. Neuroimage: Clinical 3, 155–162. https://doi.org/10.1016/j.nicl.2013.08.007.

Langston, R.F., Stevenson, C.H., Wilson, C.L., Saunders, I., Wood, E.R., 2010. The role of hippocampal subregions in memory for stimulus associations. Behav. Brain Res. 215 (2), 275–291. https://doi.org/10.1016/j.bbr.2010.07.006.

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R.A., Rueckert, D., 2018. Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. Sci. Rep. 8 (1), 1–16. https://doi.org/10.1038/s41598-018-29295-9.

Leuner, B., Gould, E., 2010. Structural plasticity and hippocampal function. Annu. Rev. Psychol. 61 https://doi.org/10.1146/annurev.psych.093008.100359, 111-C3.

Maruszak, A., Thuret, S., 2014. Why looking at the whole hippocampus is not enough—a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for Alzheimer's disease diagnosis. Front. Cell. Neurosci. 8 https://doi.org/10.3389/fncel.2014.00095.

Mcewen, B., 1999. Stress and hippocampal plasticity, 22. https://doi.org/10.1146/annurev.neuro.22.1.105.
Mueller, S.G., Yushkevich, P.A., Das, S., Wang, L., Leemput, K.V., Iglesias, J.E., et al., 2018. Systematic comparison of different techniques to measure hippocampal subfield volumes in ADNI2. Initiative, for the A. D. N Neuroimage: Clinical 17, 1006. https://doi.org/10.1016/j.nicl.2017.12.036.
Olsen, R.K., Carr, V.A., Daugherty, A.M., La Joie, R., Amaral, R.S.C., Amunts, K., et al., 2019. Progress update from the hippocampal subfields group. Alzheimer's Dementia: Diagn. Assess. Dise. Monit. 11, 439–449. https://doi.org/10.1016/j.dadm.2019.04.001.
Patel, S., Park, M.T.M., Devenyi, G.A., Patel, R., Masellis, M., Knight, J., Chakravarty, M.M., 2017. Heritability of hippocampal subfield volumes using a twin and non-twin siblings design. Hum. Brain Mapp. 38 (9), 4337–4352. https://doi.org/10.1002/hbm.23654.
Pipitone, J., Park, M.T.M., Winterburn, J., Lett, T.A., Lerch, J.P., Pruessner, J.C., et al., 2014. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. Neuroimage 101, 494–512. https://doi.org/10.1016/j.neuroimage.2014.04.054.
Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: a robust approach. Neuroimage 53 (4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020.
Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418. https://doi.org/10.1016/j.neuroimage.2012.02.084.
Sankar, T., Park, M.T.M., Jawa, T., Patel, R., Bhagwat, N., Voineskos, A.N., et al., 2017. Your algorithm might think the hippocampus grows in Alzheimer's disease: caveats of longitudinal automated hippocampal volumetry. Hum. Brain Mapp. 38 (6), 2875–2896. https://doi.org/10.1002/hbm.23559.
Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L.M., Trojanowski, J.Q., et al., 2009. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 132 (4), 1067–1077. https://doi.org/10.1093/brain/awp007.
Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psychiatry 20 (1), 11–21. https://doi.org/10.1136/jnnp.20.1.11.
Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22 (3), 1060–1075. https://doi.org/10.1016/j.neuroimage.2004.03.032.
Ségonne, Florent, Pacheco, J., Fischl, B., 2007. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. IEEE Trans. Med. Imag. 26 (4), 518–529. https://doi.org/10.1109/TMI.2006.887364.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17 (1), 87–97. https://doi.org/10.1109/42.668698.
Small, S.A., Chawla, M.K., Buonocore, M., Rapp, P.R., Barnes, C.A., 2004. Imaging correlates of brain function in monkeys and rats isolates a hippocampal subregion differentially vulnerable to aging. Proc. Natl. Acad. Sci. U. S. A. 101 (18), 7181–7186. https://doi.org/10.1073/pnas.0400285101.
Tamnes, C.K., Bos, M.G.N., van de Kamp, F.C., Peters, S., Crone, E.A., 2018. Longitudinal development of hippocampal subregions from childhood to adulthood. Dev. Cogn. Neurosci. 30, 212–222. https://doi.org/10.1016/j.dcn.2018.03.009.
Teicher, M.H., Anderson, C.M., Polcari, A., 2012. Childhood maltreatment is associated with reduced volume in the hippocampal subfields CA3, dentate gyrus, and subiculum. Proc. Natl. Acad. Sci. 109 (9), E563–E572. https://doi.org/10.1073/pnas.1115396109.
van der Flier, W.M., van Buchem, M.A., Weverling-Rijnsburger, A.W.E., Mutsaers, E.R., Bollen, E.L.E.M., Admiraal-Behloul, F., et al., 2004. Memory complaints in patients with normal cognition are associated with smaller hippocampal volumes. J. Neurol. 251 (6), 671–675. https://doi.org/10.1007/s00415-004-0390-7.
Van Leemput, K., 2009. Encoding probabilistic brain atlases using Bayesian inference. IEEE Trans. Med. Imaging 28 (6), 822–837. https://doi.org/10.1109/TMI.2008.2010434.
Whelan, C.D., Hibar, D.P., van Velzen, L.S., Zannas, A.S., Carrillo-Roa, T., McMahon, K., et al., 2016. Heritability and reliability of automatically segmented human hippocampal formation subregions. Neuroimage 128, 125–137. https://doi.org/10.1016/j.neuroimage.2015.12.039.
Winocur, G., Wojtowicz, J.M., Sekeres, M., Snyder, J.S., Wang, S., 2006. Inhibition of neurogenesis interferes with hippocampus-dependent memory function. Hippocampus 16 (3), 296–304. https://doi.org/10.1002/hipo.20163.
Winterburn, J.L., Pruessner, J.C., Chavez, S., Schira, M.M., Lobaugh, N.J., Voineskos, A.N., Chakravarty, M.M., 2013. A novel in vivo atlas of human hippocampal subfields using high-resolution 3 T magnetic resonance imaging. Neuroimage 74, 254–265. https://doi.org/10.1016/j.neuroimage.2013.02.003.
Wisse, L.E.M., Kuijf, H.J., Honingh, A.M., Wang, H., Pluta, J.B., Das, S.R., et al., 2016. Automated hippocampal subfield segmentation at 7T MRI. AJNR. Am. J. Neuroradiol. 37 (6), 1050–1057. https://doi.org/10.3174/ajnr.A4659.
Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., 2010. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. Neuroimage 52 (1), 109–118. https://doi.org/10.1016/j.neuroimage.2010.04.006.
Worker, A., Dima, D., Combes, A., Crum, W.R., Streffer, J., Einstein, S., et al., 2018. Test-retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and Alzheimer's disease populations. Hum. Brain Mapp. 39 (4), 1743–1754. https://doi.org/10.1002/hbm.23948.
Yushkevich, P.A., Amaral, R.S.C., Augustinack, J.C., Bender, A.R., Bernstein, J.D., Boccardi, M., et al., 2015. Quantitative comparison of 21 protocols for labeling hippocampal subfields and Parahippocampal subregions in in vivo MRI: towards a harmonized segmentation protocol. Neuroimage 111, 526–541. https://doi.org/10.1016/j.neuroimage.2015.01.004.
Yushkevich, P.A., Avants, B.B., Das, S.R., Pluta, J., Altinay, M., Craige, C., 2010a. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3 Tesla MRI data. Neuroimage 50 (2), 434–445. https://doi.org/10.1016/j.neuroimage.2009.12.007.
Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., et al., 2010b. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. Neuroimage 53 (4), 1208–1224. https://doi.org/10.1016/j.neuroimage.2010.06.040.
Zeineh, M.M., Engel, S.A., Thompson, P.M., Bookheimer, S.Y., 2001. Unfolding the human hippocampus with high resolution structural and functional MRI. Anat. Rec. 265 (2), 111–120.
Zhao, W., Wang, X., Yin, C., He, M., Li, S., Han, Y., 2019. Trajectories of the hippocampal subfields atrophy in the Alzheimer's disease: a structural imaging study. Front. Neuroinf. 13 https://doi.org/10.3389/fninf.2019.00013.
Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Kaus, M.R., Haker, S.J., et al., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Acad. Radiol. 11 (2), 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8.