DOI:
10.1002/hbm.23948
10.1002/hbm.23948

Document Version
Publisher's PDF, also known as Version of record

Received: 12 July 2016 | Revised: 19 November 2017 | Accepted: 28 December 2017

DOI: 10.1002/hbm.23948

RESEARCH ARTICLE

WILEY

Test–retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and Alzheimer's disease populations

Amanda Worker[1] | Danai Dima[1] | Anna Combes[1] | William R. Crum[1] | Johannes Streffer[2] | Steven Einstein[3] | Mitul A. Mehta[1] | Gareth J. Barker[1] | Steve C. R. Williams[1] | Owen O'daly[1]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, Institute of Psychiatry, London, United Kingdom

[2]Janssen-Pharmaceutical Companies of Johnson & Johnson, Janssen Research and Development, Beerse, Belgium

[3]Janssen-Pharmaceutical Companies of Johnson & Johnson, Janssen Research and Development, Titusville, New Jersey

Correspondence
Amanda Worker, Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London, SE5 8AF, UK
Email: amanda.worker@kcl.ac.uk

Funding information
National Institute for Health Research (NIHR) Biomedical Research Centre for Mental Health at South London and Maudsley NHS Foundation Trust and [Institute of Psychiatry, Psychology and Neuroscience] King's College London (Amanda Worker, Danai Dima, Mitul Mehta and Steve Williams). National Institute for Health Research (NIHR) Infrastructure grant for the Wellcome Trust/KCL Clinical Research Facility (Owen O'Daly). NARSAD 2014 Young Investigator Award (Leichtung Family Investigator) and a Psychiatric Research Trust grant (Danai Dima). Data collection was funded by Johnson & Johnson

Abstract

The hippocampal formation is a complex brain structure that is important in cognitive processes such as memory, mood, reward processing and other executive functions. Histological and neuroimaging studies have implicated the hippocampal region in neuropsychiatric disorders as well as in neurodegenerative diseases. This highly plastic limbic region is made up of several subregions that are believed to have different functional roles. Therefore, there is a growing interest in imaging the subregions of the hippocampal formation rather than modelling the hippocampus as a homogenous structure, driving the development of new automated analysis tools. Consequently, there is a pressing need to understand the stability of the measures derived from these new techniques. In this study, an automated hippocampal subregion segmentation pipeline, released as a developmental version of Freesurfer (v6.0), was applied to T1-weighted magnetic resonance imaging (MRI) scans of 22 healthy older participants, scanned on 3 separate occasions and a separate longitudinal dataset of 40 Alzheimer's disease (AD) patients. Test–retest reliability of hippocampal subregion volumes was assessed using the intra-class correlation coefficient (ICC), percentage volume difference and percentage volume overlap (Dice). Sensitivity of the regional estimates to longitudinal change was estimated using linear mixed effects (LME) modelling. The results show that out of the 24 hippocampal subregions, 20 had ICC scores of 0.9 or higher in both samples; these regions include the molecular layer, granule cell layer of the dentate gyrus, CA1, CA3 and the subiculum (ICC $>$ 0.9), whilst the hippocampal fissure and fimbria had lower ICC scores (0.73–0.88). Furthermore, LME analysis of the independent AD dataset demonstrated sensitivity to group and individual differences in the rate of volume change over time in several hippocampal subregions (CA1, molecular layer, CA3, hippocampal tail, fissure and presubiculum). These results indicate that this automated segmentation method provides a robust method with which to measure hippocampal subregions, and may be useful in tracking disease progression and measuring the effects of pharmacological intervention.

KEYWORDS

brain structure, FreeSurfer, hippocampal subfields, hippocampus, intraclass correlation coefficient, linear mixed effects, Magnetic Resonance Imaging, test-retest reliability

1 | INTRODUCTION

The hippocampal formation is a brain region forming part of the limbic system that has been implicated in many psychiatric conditions, including major depressive disorder (MDD), schizophrenia (SCZ), post-traumatic stress disorder (PTSD) and Alzheimer's disease (AD) (Arnold, 1997; Bartsch, 2012; Chakos et al., 2005; Du et al., 2001; Kempton, Salvador, Munafo, Geddes, Simmons, Frangou, & Williams, 2011; Kühn & Gallinat, 2013; Laakso et al., 1998; Videbech & Ravnkilde, 1957). Evidence suggests that the hippocampal formation is highly plastic and sensitive to stress and is believed to have a critical role in cognitive processes such as memory formation, reward processing, fear regulation, mood and other executive functions (Andersen, Morris, Amaral, Bliss, & Okeefe, 2006; Scoviille & Milner, 1957). The cornu amnonis (regions CA3, CA2, CA1), dentate gyrus, subiculum, presubiculum, parasubiculum and entorhinal cortex constitute the hippocampal formation and much like the cerebral cortex, these subregions have connections both between one another and to other brain regions via the entorhinal cortex, making this a complex, heterogeneous structure. Limitations in magnetic resonance imaging (MRI) acquisition, resolution and segmentation have meant that *in vivo* neuroimaging studies have typically been forced to model the hippocampus as a homogenous structure. This approach has been successful in identifying the hippocampus as a region that is sensitive to disease processes and reduced hippocampal volume is evident in many neurological and psychiatric conditions (Small, Schobel, Buxton, Witter, & Barnes, 2011). Furthermore, evidence in PTSD suggests that hippocampal volume may be sensitive to pharmacological intervention (Vermetten, Vythilingam, Southwick, Charney, & Bremner, 2003).

Despite the advances made from modelling the hippocampus as a whole, rodent and primate studies suggest greater focus on hippocampal subregions may be highly informative (Malberg, 2004; Malberg, Eisch, Nestler, & Duman, et al., 2000). Several research groups have developed manual segmentation protocols for hippocampal subregion segmentation (Adler et al., 2014; Kulaga-Yoskovitz et al., 2015; Mueller et al., 2007; Wisse et al., 2012), however, inconsistencies in terms of labels used, extent of labels and label boundaries make it difficult to compare findings across research groups (Yushkevich, Amaral, et al., 2015). Although manual delineation allows for a great deal of precision, it can also be time-consuming and reliant upon expertise, which is especially troublesome in large datasets that are now frequently analysed such as Alzheimer's Disease Neuroimaging Initiative (ADNI).

Several automatic or semi-automatic techniques have been developed for application to either a T1-weighted MRI scan, or a T1-weighted plus a high resolution T2-weighted scan (Leemput et al., 2009; Mri et al., 2010; Yushkevich, Pluta, et al., 2015), relying on image intensities and probabilistic atlas for segmentation. Most recently, a pipeline released as part of the FreeSurfer package (v6.0) (Iglesias et al., 2015) and is compatible with Freesurfer v5.3 (developmental version available at https://surfer.nmr.mgh.harvard.edu/fswiki/HippocampalSubfields), offers the possibility of automated segmentation of hippocampal subregions, utilising a probabilistic atlas that has been built from manual segmentation of *in vivo* and ultra-high resolution *ex vivo* data. This method is recommended for use with a T1-weighted and high-resolution T2-weighted MRI scan, but can also be applied to a standard T1-weighted scan alone. This method utilises an atlas that closely matches the regions defined during histological investigations, and reportedly improves classification accuracy of mild cognitive impairment (MCI) and AD patients by a notable 5.9%, compared to whole hippocampus measures (Iglesias et al., 2015). While clearly promising, the reliability of these novel measures is yet to be established.

A popular measure of test–retest reliability is the intra-class correlation coefficient (ICC) (Shrout & Fleiss, 1979) that can be used to assess how consistent data are between sessions or participants. Whilst others have applied reliability metrics to the FreeSurfer segmentation and parcellation previously (Liem et al., 2015; Morey et al., 2010), Whelan and colleagues have more recently shown that using the new automated segmentation method, 12 hippocampal subregions are reliable across two sessions in the Alzheimer's Disease Neuroimaging Initiative (ADNI-2) dataset (Whelan et al., 2016). However, this study has focused on cross-sectional segmentation and the reliability of these measures has not yet been assessed for data segmented using the popular longitudinal method (Reuter, Schmansky, Rosas, & Fischl, 2012). Here we address this gap in the literature and provide the first assessment of the test–retest reliability of automated hippocampal subregion volumes using both the cross-sectional and longitudinal processing approaches in two independent datasets consisting of healthy control participants and AD patients. Furthermore, we use data that has been acquired over three scanning sessions rather than two. The inclusion of an AD sample where there is likely to be greater within subject and between subject variation; as the segmentation of hippocampal subregions may be particularly relevant to research in this population. We have also included metrics for percentage volume difference and percentage volume overlap; the latter is particularly interesting as it provides some information about the variation in shape of the segmented regions. Finally, we assess the sensitivity of hippocampal subregion volume to detect longitudinal change using a linear mixed effects model (Verbeke, 1997) . The primary aim of the study was to provide metrics on the between-session reliability of automated hippocampal subregion segmentation on standard T1-weighted MRI data using ICC, percentage volume difference and percentage volume overlap (Dice).

2 | METHODS AND MATERIALS

2.1 | Healthy control cohort

Twenty-four healthy right-handed older adults aged 50–73 were recruited (M = 13, F = 11). Participants were cognitively healthy with no history of psychiatric disorder, neurological disease or taking psychoactive treatments such as antidepressants.

Participants visited the centre on three separate occasions having a baseline scan and 1-week and 4-week follow-ups.

The study was approved by the King's College London Psychiatry, Nursing and Midwifery Research Ethics subcommittee. Written informed consent was given by all of the participants before taking part in the study.

2.1.1 | Image acquisition

T1-weighted IR-SPGR 3-dimensional images were acquired from the whole brain following the ADNI GO protocol (http://adni.loni.usc.edu) on a 3T Discovery MR750 MRI scanner (General Electric, Milwaukee, USA) fitted with a standard GE head-neck-spine array, which provides 12 coil coverage of the head. Sequence parameters were repetition time (TR) = 7 ms; echo time (TE) = 3 ms; inversion time (TI) = 400 ms; flip angle 11°; 256 × 256 acquisition matrix over 270 mm field of view (FoV) yielding a 1.05 mm in plane voxel size. Sagittal slices (partitions) of thickness 196 1.2 mm were collected, giving full brain coverage. Scan time for this sequence was approximately 6.5 min.

2.2 | MIRIAD cohort

An additional longitudinal AD and age-matched healthy control sample was also included in the study. These data were collected as part of the MIRIAD longitudinal study, designed to investigate the feasibility of using MRI as an outcome measure for clinical trials in AD treatments, further details on the study can be found in the original publication (Malone et al., 2013). Participants were scanned at intervals from 2 weeks to 2 years. For the reliability analyses, the data used included the first scans of the baseline, 2 week and 6 week follow-ups, whereas linear mixed-effects modelling also included data acquired at the 12 months, 18 months and 24 months follow-ups. Data from a total of 40 AD patients were included in this study.

All images were acquired on a single 1.5T scanner (GE Signa, GE Medical Systems, Milwaukee, Wisconsin) from 2000 to 2003. Volumetric T1-weighted imaged were acquired with an IR-FSPGR (inversion recovery prepared fast spoiled gradient recalled) sequence, field of view 24 cm, 256 x 256 matrix, 124 1.5 mm slices in coronal orientation, TR 15 ms, TE 5,4 ms, flip angle 15°, and T1 650 ms.

2.3 | Image analysis (pre-processing)

2.3.1 | Standard FreeSurfer analysis pipeline

All T1-weighted images were visually inspected for motion artefact, wrap-around and grey/white contrast; it was not necessary to exclude any data. Automated whole brain segmentation and cortical reconstruction was carried out using FreeSurfer v5.3.0 (Massachusetts General Hospital, Harvard Medical School; http://surfer.nmr.mgh.harvard.edu). These well-validated and fully automated procedures have been described in detail elsewhere (Fischl et al., 2002). In brief, T1-weighted scans undergo an affine registration to MNI305 space and skull stripping. This is followed by labeling of volumetric structures based on normalised intensity and neighbour constraints. The subcortical volume segmentation procedure is completed with a high dimensional non-linear volumetric alignment to the MNI305 atlas. This cross-sectional pipeline is hereafter referred to as 'Cross'.

2.3.2 | Longitudinal pipeline

Often longitudinal datasets can contain random within-subject variation from both acquisition and processing procedures, but using a longitudinal-specific approach can significantly reduce this variability and avoids the bias associated with common approaches, such as registering all volumes to each subject's baseline scan. Longitudinal processing was carried out using FreeSurfer v5.3.0 (Reuter et al., 2012). First, a within-subject template was created for each subject using a robust, inverse consistent, registration, containing common information from each timepoint. Each timepoint was then initialised to this within-subject template, which includes normalisation (affine registration to the within-subject template) and non-linear atlas registration (same parameters applied to data from all timepoints) and then segmentation using an intensity based probabilistic voting scheme, which is driven by all timepoints initial cross-sectional segmentation thus improving reliability and statistical power. This method (hereafter referred to as 'Long') avoids processing bias by treating each timepoint in the same way, independent of order. These processing steps were repeated with the inclusion of data from two timepoints (0 and 1 week in HC and 0 and 2 weeks in AD) and three timepoints.

2.3.3 | Hippocampal subregions

A pipeline for hippocampal subregion segmentation, which has been released as part of FreeSurfer v6.0 and is compatible with Freesurfer v5.3, was applied to the soft, probabilistic segmentations of the hippocampus produced by the Cross pipeline, yielding volumetric measures of each subregion. Longitudinal hippocampal subfield segmentations were estimated using a dedicated hippocampal subfield segmentation algorithm that was applied to the subject-specific template produced during the longitudinal pipeline as detailed above (Iglesias et al., 2016).

Individual subregions were defined using a Bayesian inference approach, based on a probabilistic atlas and observed image intensities. The whole brain segmentation was used to improve the estimate of Gaussian parameters (tissue class). The major difference between the present and previous versions of Freesurfer is the probabilistic atlas, derived from manual segmentation of *in vivo* and ultra-high resolution *ex vivo* data to improve labelling (Iglesias et al., 2015). See Figure 1 for visualisation of hippocampal subregion segmentation.

2.3.4 | Quality control and exclusion criteria

After completion of the segmentation pipeline, all volumes were visually inspected and no manual edits were necessary. Volumes were also assessed so that outliers could be identified, although it was not necessary to exclude any datasets based on these measures. Two participants were excluded due to having data available from only two scans. A total of 22 participants (M = 11, F = 10) were included in the final statistical analysis.

2.4 | Statistical methods

2.4.1 | Test-retest reliability

To evaluate the test–retest reliability of automated hippocampal subregion measures, we examined the inter-session variability of volumetric measures. The third form of the Intraclass Correlation Coefficient ($ICC_{3,1}$), as defined by Shrout and Fleiss (1979), was calculated for each

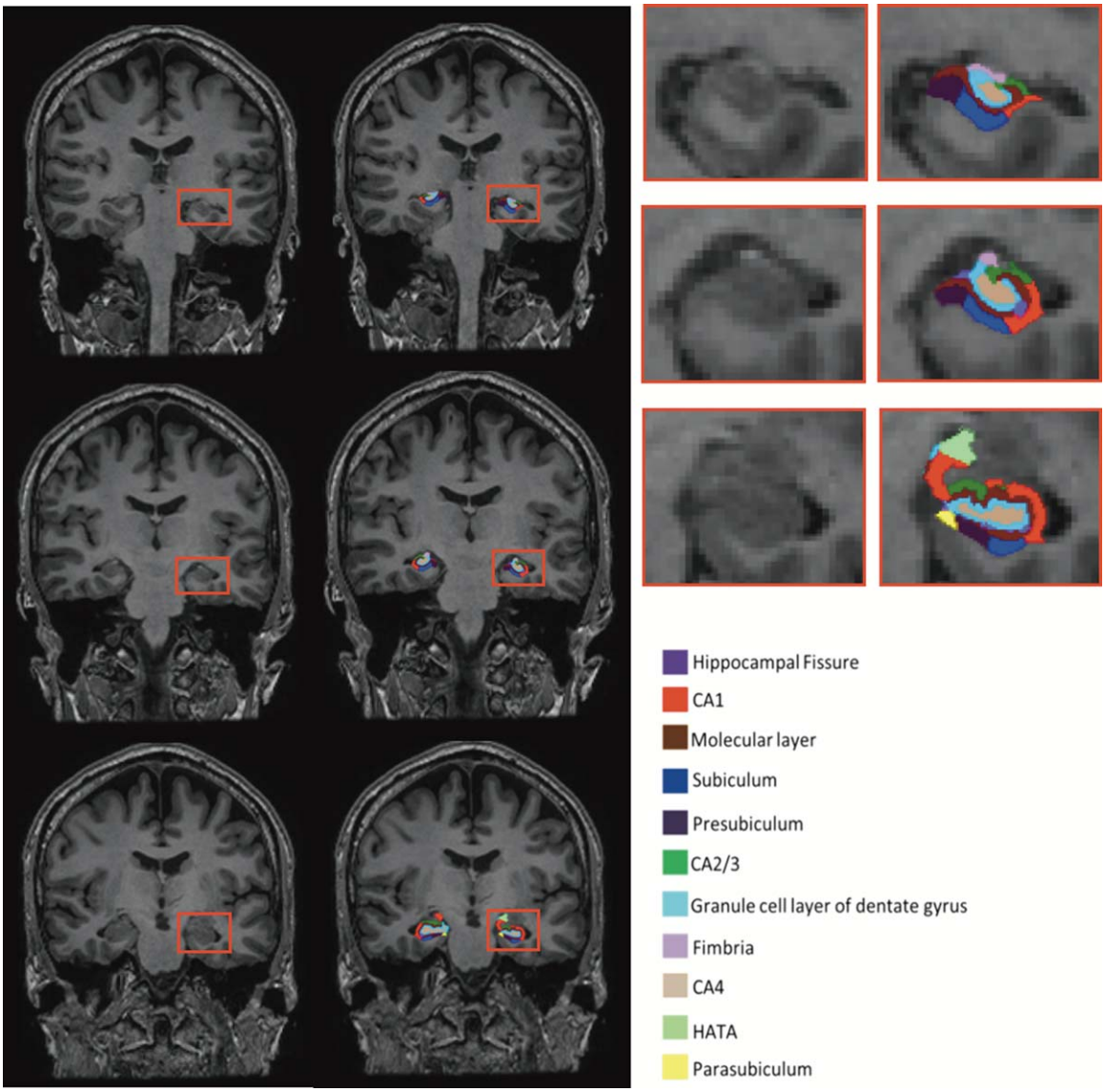

FIGURE 1 Visualisation of hippocampal subregion segmentation of a single subject. Regions not visible in this view are: alveus, hippocampal fissure and hippocampal tail

region of interest, estimating the correlation of measures between the three sessions. This was then repeated to estimate the correlation of measures between two sessions. The ICC was modelled by a two-way mixed effects model; random subject effects and fixed sessions effects, with absolute agreement. A statistical toolbox designed for ICC analysis (Caceres, Hall, Zelaya, Williams, & Mehta, 2009) implemented using Matlab 8.0.0, was used to calculate ICC_{max} values from the mean volumes of predefined regions.

Percent volume difference is given by Equation (1) where an optimal value of zero is achieved for identical volumes and an increase in values indicating greater volume difference. Percent volume overlap is given by Equation (2) where an optimal value of zero is achieved for identical volumes and an increase in values indicating greater volume difference. Percent volume difference and percent volume overlap were calculated using Freesurfer **mri_compute_overlap** function. In these equations, A points to the volume measure of timepoint A and B points to the volume measure of timepoint B, these timepoints are substituted to calculate the volume differences between timepoints A vs B, A vs C and B vs C. Percent volume difference was also calculated for the baseline and 1 year follow-up scan for all available data from the MIRIAD cohort.

$$\text{Volume difference} = \frac{2 * (|A| - |B|)}{|A| + |B|} \times 100 \quad (1)$$

$$\text{Volume overlap} = \frac{2 * |A| \cap |B|}{|A| + |B|} \times 100 \quad (2)$$

Studies previously assessing the test–retest reliability of automated segmentation of brain regions have differed in the methods adopted, with some reporting test–retest reliability of segmentations produced by cross-sectional methods (Morey et al., 2010; Whelan et al., 2016) whilst others report on the segmentation produced by

TABLE 1 Intraclass correlation coefficient for hippocampal subregion volumes—two timepoints

		Healthy						AD					
		Cross			Long			Cross			Long		
Region		ICC	CI Lower	CI Upper	ICC	CI Lower	CI Upper	ICC	CI Lower	CI Upper	ICC	CI Lower	CI Upper
Hippocampal Tail	Left	0.90	0.78	0.96	0.97*	0.93	0.99	0.92	0.85	0.96	0.98*	0.96	0.99
	Right	0.90	0.78	0.96	0.96	0.91	0.98	0.94	0.88	0.97	0.97	0.94	0.99
Subiculum	Left	0.91	0.79	0.96	0.98*	0.94	0.99	0.96	0.93	0.98	0.98	0.97	0.99
	Right	0.92	0.81	0.96	0.96	0.91	0.98	0.95	0.90	0.97	0.99*	0.98	0.99
CA1	Left	0.91	0.80	0.96	0.97*	0.94	0.99	0.95	0.90	0.97	0.99*	0.97	0.99
	Right	0.87	0.72	0.95	0.98*	0.95	0.99	0.91	0.84	0.95	0.99*	0.98	0.99
Fissure	Left	0.67	0.36	0.85	0.85	0.68	0.94	0.94	0.90	0.96	0.95	0.91	0.98
	Right	0.70	0.40	0.86	0.82	0.62	0.92	0.88	0.79	0.94	0.91	0.84	0.96
Presubiculum	Left	0.90	0.77	0.96	0.97*	0.92	0.99	0.92	0.86	0.85	0.97*	0.94	0.98
	Right	0.79	0.55	0.91	0.96*	0.90	0.98	0.89	0.80	0.94	0.97*	0.94	0.98
Parasubiculum	Left	0.87	0.72	0.95	0.96*	0.90	0.98	0.74	0.56	0.85	0.98*	0.95	0.99
	Right	0.70	0.37	0.85	0.96*	0.90	0.98	0.70	0.50	0.83	0.95*	0.91	0.98
Molecular Layer	Left	0.95	0.89	0.98	0.98	0.96	0.99	0.97	0.94	0.98	0.99*	0.98	0.99
	Right	0.90	0.77	0.96	0.99*	0.97	0.99	0.95	0.91	0.97	0.99*	0.99	0.99
GC-DG	Left	0.91	0.79	0.96	0.97*	0.93	0.99	0.97	0.94	0.98	0.99*	0.98	0.99
	Right	0.78	0.54	0.90	0.98*	0.95	0.99	0.95	0.91	0.97	0.99*	0.99	0.99
CA3	Left	0.73	0.45	0.88	0.96*	0.90	0.98	0.94	0.89	0.97	0.97	0.94	0.98
	Right	0.78	0.55	0.91	0.97*	0.93	0.99	0.93	0.88	0.96	0.99*	0.98	0.99
CA4	Left	0.89	0.76	0.95	0.97*	0.93	0.99	0.97	0.94	0.98	0.99*	0.97	0.99
	Right	0.75	0.49	0.89	0.97*	0.92	0.99	0.95	0.90	0.97	0.99*	0.98	0.99
Fimbria	Left	0.89	0.75	0.95	0.97*	0.92	0.99	0.62	0.39	0.78	0.80*	0.64	0.89
	Right	0.86	0.69	0.94	0.94	0.87	0.98	0.76	0.58	0.87	0.90*	0.82	0.95
HATA	Left	0.81	0.60	0.92	0.86	0.69	0.94	0.87	0.77	0.93	0.97*	0.95	0.99
	Right	0.60	0.25	0.81	0.92*	0.83	0.97	0.84	0.71	0.91	0.98*	0.96	0.99
Whole	Left	0.96	0.91	0.98	0.99*	0.97	0.99	0.97	0.94	0.98	0.99*	0.98	0.99
	Right	0.92	0.83	0.97	0.99*	0.98	0.99	0.97	0.93	0.98	0.99*	0.99	0.99

GC-DG (Dentate Gyrus Granule Cell Layer), HATA (hippocampal-amygdaloid transition area). Whole Hippocampus represents the measure of hippocampal volume produced by the hippocampal subregion segmentation pipeline. *Significantly different based on point estimate of 'long' not lying within the confidence interval of 'cross'.

longitudinal processing (Liem et al., 2015; Morey et al., 2010). Thus we report reliability metrics for both Cross and Long processing streams.

2.4.2 | Linear mixed effects model

Data from the MIRIAD cohort, spanning 2 years, was included in a linear mixed effects model to assess longitudinal change, with fixed effects of group (AD or HC) and two random effects of intercept and slope. Hippocampal subregion volumes were first corrected for intracranial volume (ICV) by dividing each volume by the ICV of that subject. Age and gender were included as nuisance variables. The interaction term (group × time) was tested to give estimates of longitudinal change by group. The model was repeated for all available data spanning the first 6 weeks to validate the observed rates of change. All analyses were performed in Matlab R2012b.

3 | RESULTS

3.1 | Participants

Twenty-two healthy control participants were included in analysis with a mean age of 59 (50–73 years). Forty AD patients were included, with a mean age of 70 (55–86 years).

TABLE 2 Intraclass correlation coefficient for hippocampal subregion volumes

		Healthy						AD					
		Cross			Long			Cross			Long		
Region		ICC	CI Lower	CI Upper	ICC	CI Lower	CI Upper	ICC	CI Lower	CI Upper	ICC	CI Lower	CI Upper
Hippocampal Tail	Left	0.88	0.81	0.96	0.98*	0.96	0.99	0.92	0.89	0.97	0.97	0.96	0.99
	Right	0.82	0.71	0.94	0.96*	0.94	0.99	0.94	0.93	0.98	0.98	0.97	0.99
Subiculum	Left	0.90	0.85	0.97	0.98*	0.96	0.99	0.94	0.93	0.98	0.99*	0.98	0.99
	Right	0.89	0.82	0.97	0.97	0.95	0.99	0.96	0.94	0.98	0.99*	0.98	0.99
CA1	Left	0.92	0.88	0.98	0.98	0.97	0.99	0.94	0.92	0.98	0.99*	0.98	0.99
	Right	0.88	0.81	0.97	0.99*	0.98	0.99	0.93	0.91	0.98	0.99*	0.99	0.99
Fissure	Left	0.66	0.47	0.88	0.79	0.67	0.94	0.94	0.92	0.98	0.97	0.96	0.99
	Right	0.64	0.45	0.88	0.73	0.58	0.91	0.87	0.82	0.95	0.93	0.92	0.98
Presubiculum	Left	0.91	0.85	0.97	0.96	0.93	0.99	0.87	0.83	0.95	0.97*	0.97	0.99
	Right	0.83	0.74	0.95	0.96*	0.94	0.99	0.89	0.86	0.96	0.97*	0.96	0.99
Parasubiculum	Left	0.84	0.75	0.95	0.95	0.92	0.98	0.72	0.63	0.89	0.97*	0.96	0.99
	Right	0.77	0.65	0.93	0.97*	0.96	0.99	0.71	0.62	0.89	0.96*	0.94	0.98
Molecular Layer	Left	0.96	0.93	0.99	0.99	0.98	0.99	0.96	0.95	0.99	0.99	0.98	0.99
	Right	0.89	0.83	0.97	0.99*	0.98	0.99	0.96	0.95	0.99	0.99	0.99	0.99
GC-DG	Left	0.90	0.85	0.97	0.98*	0.96	0.99	0.95	0.94	0.98	0.99*	0.99	0.99
	Right	0.81	0.70	0.94	0.98*	0.97	0.99	0.95	0.93	0.98	0.99*	0.99	0.99
CA3	Left	0.77	0.64	0.93	0.97*	0.96	0.99	0.94	0.92	0.98	0.98	0.97	0.99
	Right	0.83	0.73	0.95	0.98*	0.97	0.99	0.94	0.92	0.98	0.98	0.98	0.99
CA4	Left	0.88	0.82	0.97	0.97	0.96	0.99	0.95	0.93	0.98	0.99*	0.98	0.99
	Right	0.78	0.66	0.93	0.98*	0.97	0.99	0.94	0.92	0.98	0.99*	0.99	0.99
Fimbria	Left	0.88	0.81	0.96	0.96	0.94	0.99	0.72	0.63	0.89	0.81	0.76	0.93
	Right	0.86	0.78	0.99	0.94	0.90	0.98	0.77	0.70	0.91	0.88	0.86	0.96
HATA	Left	0.79	0.67	0.93	0.94*	0.90	0.98	0.84	0.79	0.94	0.96*	0.95	0.99
	Right	0.67	0.49	0.89	0.91*	0.86	0.97	0.88	0.84	0.96	0.96	0.95	0.99
Whole	Left	0.96	0.94	0.99	0.99	0.98	0.99	0.96	0.94	0.99	0.99	0.99	0.99
	Right	0.91	0.86	0.97	0.99*	0.99	0.99	0.97	0.96	0.99	0.99	0.99	0.99

GC-DG (Dentate Gyrus Granule Cell Layer), HATA (hippocampal-amygdaloid transition area). Whole Hippocampus represents the measure of hippocampal volume produced by the hippocampal subregion segmentation pipeline. *Significantly different based on point estimate of 'long' not lying within the confidence interval of 'cross'.

3.2 | Test–retest reliability

The test–retest reliability of hippocampal subregion volumes varied between regions (Tables 1 and 2). With the inclusion of two timepoints, all regions with the exception of the fissure and left HATA achieve ICC scores of > 0.9 in the HC sample, whilst all regions achieve ICC scores > 0.9 in the AD sample. ICC scores of the volumes from the longitudinal stream of Freesurfer were significantly higher than for the cross-sectional stream in the CA1, presubiculum, parasubiculum, granule cell layer of the dentate gyrus, CA3, CA4, whole hippocampus, left hippocampal tail, subiculum, fimbria and right HATA. In AD, the results were similar with the addition of the right CA1 and fimbria, and left molecular layer and HATA (see Table 1).

With the inclusion of three timepoints, all regions with the exception of the hippocampal fissure achieve ICC scores > 0.9 in the HC sample, whilst all regions with the exception of the fimbria achieve ICC scores > 0.9 in the AD sample over a 6 week period. ICC scores of the volumes from the longitudinal stream of Freesurfer were significantly higher than for the cross-sectional stream in the hippocampal tail, granule cell layer of the dentate gyrus, CA3, HATA, left subiculum, right CA1, presubiculum, parasubiculum, molecular layer, CA4 and whole hippocampus in the HC group.

Longitudinal processing yielded significantly higher ICC scores than cross-sectional processing in the subiculum, CA1, presubiculum, parasubiculum, granule cell layer of the dentate gyrus, CA4 and left HATA (see Table 2).

Figure 2 displays the percent volume difference between repeated scanning sessions values represent the mean volume difference for each session comparison (A vs B, A vs C, B vs C). The molecular layer, CA1, granule cell layer of the dentate gyrus and whole hippocampus show the most consistency in size across timepoints, whilst the fimbria, hippocampal fissure and parasubiculum show the least consistency.

Figure 3 displays the percent volume overlap(Dice) between repeated scanning sessions, values represent the mean volume overlap for each session comparison (A vs B, A vs C, B vs C). The greatest overlap was detected in the whole hippocampus and CA4, and the least overlap in the hippocampal fissure.

3.3 | Linear mixed effects model

The results from the linear mixed effects model are summarised in Table 3. A significant interaction of group × time was found bilaterally in the whole hippocampus and in the right CA1, CA3, molecular layer and left presubiculum and hippocampal tail, over a 2-year period. Importantly, over the 6-week period, no significant volume differences were evident.

It should be noted that our findings are in general agreement with a recently published paper with similar aims (Iglesias et al., 2016). Iglesias and colleagues similarly found that volumetric estimates from hippocampal subfields were reliable (volume difference and overlap) and sensitive to AD-related decline. However, we present here additional information with ICC scores, percentage volume difference/overlap from scans taken weeks apart rather than same day scans and also reliability metrics for a completely independent dataset of healthy control participants.

4 | DISCUSSION

In this study, we have assessed the test–retest reliability of automated hippocampal subregion segmentation of standard, ADNI-compatible; T1-weighted MRI scans. The results from this study show that almost all hippocampal subregion segmentations achieve high ICC scores (ICC $>$ 0.85), after longitudinal processing compared to just over half after cross-sectional processing. To our knowledge this is the first study to assess the test–retest reliability using ICC of automated hippocampal subregion segmentation applied to cross-sectional and longitudinal data using Freesurfer's pipeline, in two independent datasets consisting of three separate timepoints, spanning 4 weeks in healthy controls and 6 weeks in AD patients.

Our results show that almost all regions are highly stable, whilst the fissure is the least stable in healthy control participants. These results are broadly in line with those of Whelan et al with slightly greater reliability in the most stable regions. The hippocampal fissure is a vestigial space located between the molecular layer and the dentate gyrus; the boundary between the fissure and extrahippocampal cerebrospinal fluid (CSF) may contribute to the lower test–retest reliability, in addition to the small size and shape which may make this region more susceptible to partial volume effects. Other regions with slightly lower test–retest reliability are the parasubiculum, fimbria and HATA which are among the smallest of hippocampal subregions.

The position of the hippocampus lying close to the skull and inferior ventricles make this area vulnerable to image distortion, artefacts and signal dropout. This, in combination with the limited resolution of T1-weighted data for subcortical structures means that the internal boundaries of the hippocampus are not visible, thus it is likely that the anatomical priors of the atlas become heavily relied upon. However, it should be noted that while the volumes measured appear to be highly reliable, the overlap measures were less consistent across the regions. In healthy controls, the longitudinal pipeline provides overlap scores in many regions that approach that seen for the whole hippocampus but importantly the cross-sectional segmentation provides poorer levels of overlap. Unsurprisingly, for data acquired over a 6-week period, in the AD patients regional overlap is poorer across sessions although perhaps surprisingly ICC scores remain high. Together, these findings suggest that there are additional factors that need to be addressed. Some of these issues could arise as a consequence of partial volume effects, one could address these issues by using $<$1 mm voxels (Ekstrom et al., 2009) and/or including an additional T2-weighted or Proton Density volume in the processing stream (Iglesias et al., 2015). Despite these limitations, compared to previous versions (Leemput et al., 2009), the methods described by Iglesias provide an improved atlas with which to define subregion boundaries that is likely to provide additional information regarding individual subregions. Indeed, Iglesias and colleagues have shown that classification of MCI and AD patients improved by 5.9% when using hippocampal subregions volume over the standard whole hippocampal volume produced by the FreeSurfer pipeline.

The results presented here reflect test–retest reliability of measures estimated from a T1-weighted scan, however it is important to emphasise that the algorithm has not yet been validated against manual segmentations and in the present study we have not assessed the accuracy of segmented regions. In future studies, a comparison will need to be made between automated and manual segmentation. Manual segmentation of hippocampal subregions itself is a vast area of research (Adler et al., 2014; Kulaga-Yoskovitz et al., 2015; Mueller et al., 2007; Wisse et al., 2012), where there is great variation in the size, shape and position of labels used (Yushkevich, Amaral, et al., 2015) making it difficult to compare automated measures with existing manual delineations. Yuskevich and colleagues have identified major areas of disagreement and taken steps towards a universally agreed method of labelling the hippocampal formation. As a result of this ambiguity, any inferences drawn from our findings are likely applicable only to datasets acquired with the same image parameters, processing pipeline and anatomical atlas as described in the methods section of this article.

This study offers several additional and informative results over the previously published reliability work of (Whelan et al., 2016), with the inclusion of volume overlap and volume difference metrics along with a longitudinal model of volume change in AD compared to HC. Furthermore, the ICC scores reported here reflect test–retest reliability of data collected across three timepoints on separate days as well as

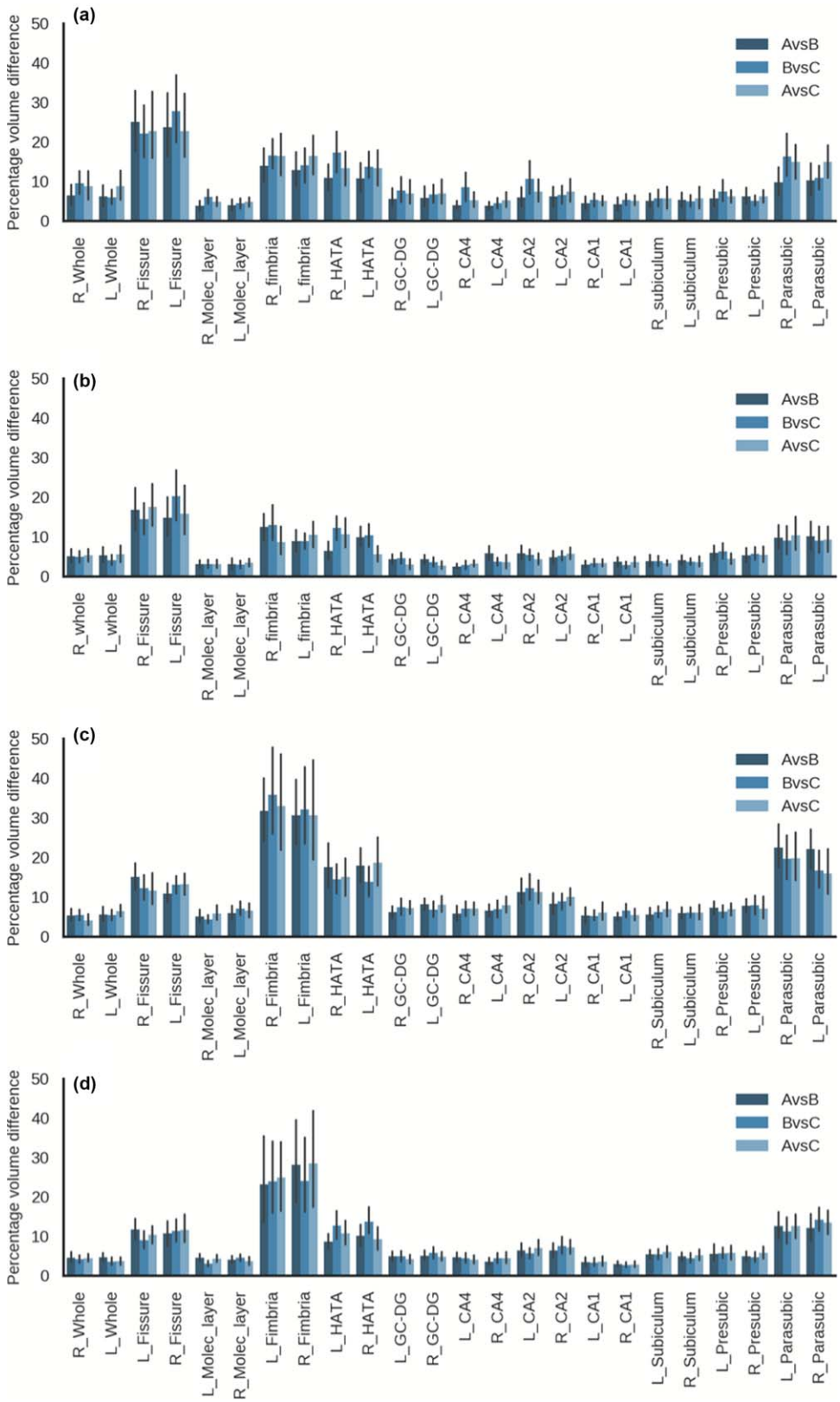

FIGURE 2 Percentage volume difference means and 95% confidence intervals. (a) Cross (b) Long, healthy control participants scanned at baseline (A), 1 week (B) and 4 weeks (C). (c) Cross (d) Long, AD sample scanned at baseline (A), 2 weeks (B) and 6 weeks (C). Whole (whole hippocampus), Fissure (hippocampal fissure), molec layer (molecular layer), HATA (hippocampal-amygdaloid transition area), GC-DG (Granule cell layer of the dentate gyrus), presubic (presubiculum) and parasubic (parasubiculum) [Color figure can be viewed at wileyonlinelibrary.com]

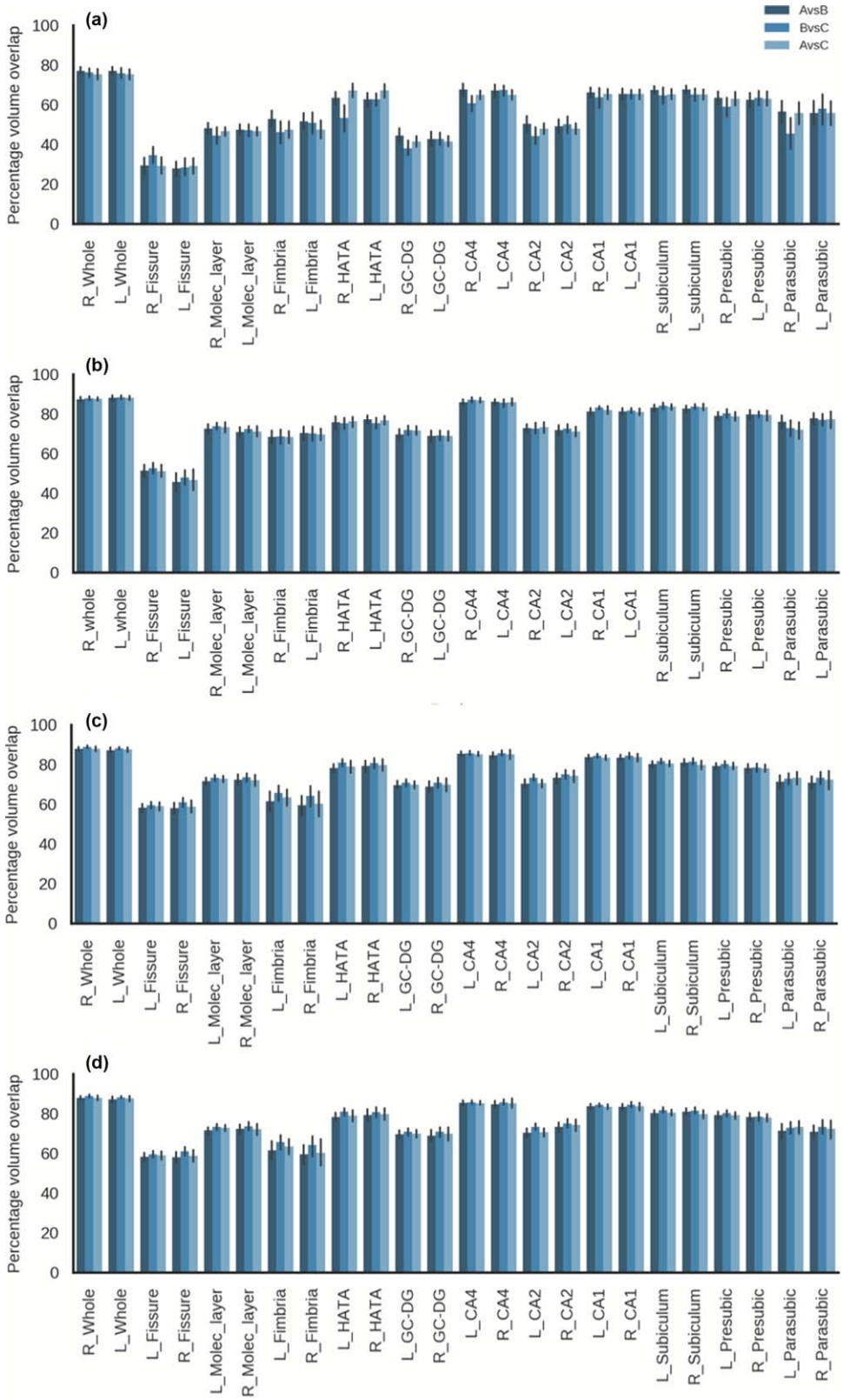

FIGURE 3 Percentage volume overlap means and 95% confidence intervals. Top: (a) Cross (b) Long, healthy control participants scanned at baseline (A), 1 week (B) and 4 weeks (C). Bottom: (a) Cross (b) Long, AD sample scanned at baseline (A), 2 weeks (B) and 6 weeks (C). Whole (whole hippocampus), Fissure (hippocampal fissure), molec layer (molecular layer), HATA (hippocampal-amygdaloid transition area), GC-DG (Granule cell layer of the dentate gyrus), presubic (presubiculum), parasubic (parasubiculum) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Results from linear mixed effects model of longitudinal change in volume in AD compared to HC

		2 years		6 weeks	
Region		***p*-value**	***F***	***p*-value**	***F***
Tail	Left	0.02*	5.71	0.19	1.75
	Right	0.05	4.14	0.26	1.31
Subiculum	Left	0.07	3.43	0.85	0.04
	Right	0.05	4.25	0.14	2.23
CA1	Left	0.33	0.95	0.94	0.01
	Right	0.002**	11.49	0.24	1.39
Fissure	Left	0.59	0.29	0.24	1.42
	Right	0.02*	6.02	0.40	0.71
Presubiculum	Left	0.01*	7.59	0.43	0.64
	Right	0.13	2.40	0.79	0.07
Parasubiculum	Left	0.21	1.62	0.45	0.57
	Right	0.28	1.19	0.18	1.84
Molecular Layer	Left	0.06	3.69	0.95	0.004
	Right	0.004**	9.05	0.53	0.40
GC-DG	Left	0.08	3.17	0.92	0.01
	Right	0.38	0.79	0.56	0.34
CA3	Left	0.39	0.76	0.77	0.09
	Right	0.01*	8.24	0.95	0.01
CA4	Left	0.07	3.35	0.97	0.001
	Right	0.45	0.58	0.69	0.16
Fimbria	Left	0.62	0.26	0.68	0.17
	Right	0.75	0.10	0.95	0.01
HATA	Left	0.45	0.59	0.80	0.06
	Right	0.33	60.96	0.99	0.00
Whole	Left	0.03*	5.20	0.65	0.20
	Right	0.01*	7.88	0.89	0.02

GC-DG (Dentate Gyrus Granule Cell Layer), HATA (hippocampal-amygdaloid transition area). Whole Hippocampus represents the measure of hippocampal volume produced by the hippocampal subregion segmentation pipeline. *<0.05 **<0.01.

two timepoints; Whelan and colleagues included data from two scanning sessions. Despite the smaller sample size of this study in comparison to those mentioned previously, measuring the reliability over a greater number of samples would typically be expected to result (via regression to the mean) in an ICC estimate closer to the true mean reliability. While our data suggest that, for the longitudinal pipeline in healthy controls, such ICC differences are small when using one additional scan for the longitudinal pipeline, it should be noted that the confidence intervals on those estimates tended to be smaller. For the cross-sectional pipeline, the ICC values differed more for two timpoint and 3 timepoint ICC calculations, but not systematically. The confidence intervals for the data processed with the cross pipeline also tended towards being tighter when the ICC was calculated over 3 data-points. Finally, a similar pattern was also evident within the AD sample.

While overall, we found that the difference between using two or three timepoints to be modest, there is a growing trend in trials to use longitudinal designs, rather than simple cross-over protocols, and a general growth in large-scale prospective imaging studies. Consequently, reliability estimates over more than two scanning sessions should results in greater confidence in stability of the data and our estimates of that estimate. Due to the small sample sizes included in the present study, we suggest that it would be beneficial in future to assess these reliability metrics in a larger sample with three or more timepoints to validate our findings.

Methods specifically developed for the processing of longitudinal structural MRI have been applied to the data in this study (Reuter et al., 2012), differing from the approach used by Whelan and colleagues (Whelan et al., 2016). Longitudinal data analysis is often limited by random variation in the data that is due to anatomical variations, acquisition procedures, for example a change of head position between scans, and processing variations associated with automated segmentation algorithms. The longitudinal processing pipeline offers a method for reducing the random variation that may arise because of processing procedures and avoids (order-based) resampling bias commonly seen with analysis of longitudinal data. Furthermore, initialising segmentation with a within-subject template simply provides a starting point only, the segmentation evolves freely for each timepoint, allowing variation in the data to be reflected in the volume estimation. Whilst our findings suggest that the using the longitudinal pipeline reduces some proportion of the random variation in the data, factors such as head motion and position, hydration of the participants and scanner instabilities are still likely to contribute to between-session variance. This method is therefore ideal for use on data in this study whereby scans were obtained on three different days making variation in the data more likely than scans obtained in the same scanning session. Our results show that using the longitudinal pipeline, ICC scores obtained from three scanning sessions show test–retest reliability scores comparable to those achieved with only two-sessions (Whelan et al., 2016) and when data was acquired on the same day (Liem et al., 2015) demonstrating the ability of this method to deal with random variation without compromising reliability.

Finally, the inclusion of an AD sample in this study offers added information on the applications of these methods in samples that are likely to have pathology of the hippocampus leading to greater between and within subject variability. This atrophy is also likely to challenge the anatomical priors of an atlas that has been built on a healthy control sample. The results reported here show that even in an AD sample the measures are stable across time, which further emphasises the need to validate these measures against manual delineation.

Results from our linear mixed effects model show that in AD there is significantly greater atrophy over a two year period in hippocampal subregions that have previously been identified by manual delineation and histological examination, such as CA1 and whole hippocampus (Mueller et al., 2010; Simic, Kostovic, Winblad, & Bogdanovic, 1997; Wisse et al., 2014). We also find greater volume loss in the molecular layer, previously associated with MCI subjects (Iglesias et al., 2015),

CA3, hippocampal tail, presubiculum and HATA and a trend towards volume loss in the subiculum, granule cell layer of the dentate gyrus and CA4. Whilst our findings generally support those published previously (Iglesias et al., 2016) we find more modest results that do not fully replicate findings using the same technique and this is likely to be due to variation in processing methods and software. Ultimately, our results indicate that firstly, the hippocampal subregion segmentation produces volume estimate which are stable over a short period, but this method is also sensitive to biologically plausible rates of volume change over time in a region specific manner (CA1), which accords with previously published literature. However, we must interpret these results with caution as it is possible that methodological issues are contributing to this effect. The hippocampus is a region that is sensitive to motion artefact and signal dropout and it is possible that the AD group is more susceptible to this; therefore, we cannot conclude that this is a true biological effect without further investigation.

Measuring the individual subregions of the hippocampus has typically been difficult due to the limited spatial resolution in human MRI, therefore neuroimaging studies have relied on measuring the hippocampus as one structure (Chupin et al., 2009; Kempton, 2011; Videbech & Ravnkilde, 1957; Erickson, Voss, Shaurya, Basak, & Szabo, 2011), or manual tracing of subregions (Kulaga-Yoskovitz et al., 2015; Mueller et al., 2007; Wisse et al., 2012). Treating the hippocampus as one structure may mean that crucial information is missed, while manual tracing of the structure can be time consuming and subjective. Thus reliable, automated segmentation of the hippocampal subregions is advantageous and has many potential applications in psychiatry and neurology. The dentate gyrus is one of the few regions where neurogenesis is known to continue into adulthood in humans (Eriksson et al., 1998) and animal studies show that therapies such as exercise and drug treatment can actually promote neurogenesis (Ho, Hooker, Sahay, Holt, & Roffman, 2013; Malberg et al., 2000) and that this is in line with the time frame of therapeutic effect in humans (Duman, Heninger, & Nestler, 2017). There is currently no definitive measure of neurogenesis *in vivo* in humans, but neuroimaging offers the possibility of bridging the gap, by using structural and functional indices as a proxy measure of neurogenesis. Vermetten et al. (2003) report a 4.6% increase in whole hippocampal volume in PTSD patients after treatment with a selective serotonin reuptake inhibitor (SSRI); this effect was seen over 36–48 weeks of drug treatment in a relatively small sample of twenty three patients. It would be of great interest if this effect could be definitively localised to the dentate gyrus.

In conclusion, we have presented test–retest reliability of automated hippocampal subregion measures in two independent longitudinal datasets of healthy older participants and AD patients. Using ICC, volume difference and volume overlap measures we have been able to quantify the reliability of hippocampal subregion volumes showing that most regions have high test–retest reliability and using linear mixed effects model we show that these measures are sensitive enough to detect change over time where it would be expected. These results indicate that the methods applied are stable and have the potential to be used as a marker of disease progression, as well as to assess the effects of pharmacological interventions.

ACKNOWLEDGMENTS

Amanda Worker, Danai Dima, Mitul Mehta and Steve Williams are partially funded by the National Institute for Health Research (NIHR) Biomedical Research Centre for Mental Health at South London and Maudsley NHS Foundation Trust and [Institute of Psychiatry, Psychology and Neuroscience] King's College London. Owen O'Daly receives salary support from a National Institute for Health Research (NIHR) Infrastructure grant for the Wellcome Trust/KCL Clinical Research Facility. Danai Dima is partially supported by a NARSAD 2014 Young Investigator Award (Leichtung Family Investigator) and a Psychiatric Research Trust grant. Gareth Barker receives honoraria for teaching for General Electric Healthcare, and acts as a consultant for IXICO. Data collection was funded by Johnson & Johnson.

ORCID

Amanda Worker http://orcid.org/0000-0003-3103-867X

REFERENCES

Adler, D. H., Pluta, J., Kadivar, S., Craige, C., Gee, J. C., Avants, B. B., & Yushkevich, P. A. (2014). NeuroImage histology-derived volumetric annotation of the human hippocampal subfields in postmortem MRI. *NeuroImage*, *84*, 505–523.

Andersen, P., Morris, R., Amaral, D., Bliss, T., & Okeefe, J. (2006). *The hippocampus book*, United States of America: Oxford University Press.

Arnold, S. (1997). The medial temporal lobe in schizophrenia. In *The neuropsychiatry of limbic and subcortical disorders* (pp. 155–166). Washington DC: American Psychiatric Press, Inc.

Bartsch, T. (2012). The clinical neurobiology of the hippocampus: An integrative view. (1st ed.). Spain: Oxford University Press.

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., & Mehta, M. A. (2009). NeuroImage measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage*, *45*(3), 758–768.

Chakos, M. H., Schobel, S. A., Gu, H., Gerig, G., Bradford, D., Charles, C., & Lieberman, J. A. (2005). Duration of illness and treatment effects on hippocampal volume in male patients with schizophrenia. *British Journal of Psychiatry*, *i*(186), 26–31.

Chupin, M., Gerardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehericy, S., … Colliot, O. (2009). Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data From ADNI. *Hippocampus*, *587*(19), 579–587.

Du, A. T., Schu, N. V., Amend, D., Laakso, M. P., Hsu, Y. Y., Jagust, W. J., … Street, C. (2001). Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. *J Neurol Neurosurg Psychiatry*, *1*, 441–447.

Duman, R. S., Heninger, G. R., & Nestler, E. J. (2017). A molecular and cellular theory of depression. *Archives of General Psychiatry*, *54*, 597–606.

Ekstrom, A. D., Bazih, A. J., Suthana, N. A., Al-hakim, R., Ogura, K., Zeineh, M., … Bookheimer, S. Y. (2009). NeuroImage advances in high-resolution imaging and computational unfolding of the human hippocampus. *NeuroImage*, *47*(1), 42–49.

Erickson, K. I., Voss, M. W., Shaurya, R., Basak, C., & Szabo, A. (2011). Exercise training increases size of hippocampus and improves memory. *PNAS*, *108*(7), 3017–3022.

Eriksson, P. S., Perfilieva, E., Bjork-Eriksson, T., Alborn, A.-M., Nordborg, C., Peterson, D. A., & Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine*, *4*(11), 1313–1317.

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

Ho, N., Hooker, J., Sahay, A., Holt, D., & Roffman, J. (2013). In vivo imaging of adult hippocampal neurogenesis: progress, pitfalls and promise. *Molecular Psychiatry*, *18*(4), 404–416.

Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., … Van Leemput, K. (2015). NeuroImage 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.

Iglesias, J. E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., & Reuter, M. (2016). NeuroImage Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. *NeuroImage*, *141*, 542–555.

Kempton, M. J., Salvador, Z., Munafo, M. R., Geddes, J. R., Simmons, A., Frangou, S., & Williams, S. C. R. (2011). Structural neuroimaging studies in major depressive disorder. *Archives of General Psychiatry*, *68*(7), 675–690.

Kühn, S., & Gallinat, J. (2013). Gray matter correlates of posttraumatic stress disorder : a quantitative meta-analysis. *Biological Psychiatry*, *73* (1), 70–74.

Kulaga-Yoskovitz, J., Bernhardt, B. C., Hong, S., Mansi, T., Liang, K. E., van der Kouwe, A. J. W., & Bernasconi, N. (2015). Multi-contrast submillimetric 3Tesla hippocampal sub field segmentation protocol and dataset. *Scientific Data*, *2*, 1–9.

Laakso, M. P., Soininen, H., Partanen, K., Lehtovirta, M., & Hallikainen, M. (1998). MRI of the hippocampus in Alzheimer's disease : sensitivity, specificity, and analysis of the incorrectly classified subjects. *Neurobiology of Aging*, *19*(1), 23–31.

Liem, F., Mérillat, S., Bezzola, L., Hirsiger, S., Philipp, M., Madhyastha, T., & Jäncke, L. (2015). NeuroImage reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly. *NeuroImage*, *108*, 95–109.

Malberg, J. E., Eisch, A. J., Nestler, E. J., & Duman, R. S. (2000). Chronic antidepressant treatment increases neurogenesis in adult rat hippocampus. *The Journal of Neuroscience*, *20*(24), 9104–9110.

Malberg, J. E. (2004). Implications of adult hippocampal neurogenesis in antidepressant action. *Journal of Psychiatry Neuroscience*, *29*(3), 196–205.

Malone, I. B., Cash, D., Ridgeway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C., & Schott, J. M. (2013). MIRIAD - Public release of a multiple time point Alzheimer's MR imaging dataset. *Neuroimage*, *70*, 33–36.

Morey, R. A., Selgrade, E. S., Ryan, H., Ii, W., Huettel, S. A., Wang, L., … Carolina, N. (2010). Scan – Rescan Reliability of Subcortical Brain Volumes Derived From Automated Segmentation. *Human Brain Mapping*, *1762* (May 2009), 1751–1762.

Mueller, S. G., Schuff, N., Yaffe, K., Madison, C., Miller, B., & Weiner, M. W. (2010). Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease. *1347*, 1339–1347.

Mueller, S. G., Stables, L., Du, A. T., Schuff, N., Truran, D., Cashdollar, N., & Weiner, M. W. (2007). Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4 T. *Neurobiology of Aging*, *28*, 719–726.

Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). NeuroImage within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, *61*(4), 1402–1418.

Scoviille, W., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*, 11–21.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations : uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Simic, G., Kostovic, I., Winblad, B., & Bogdanovic, N. (1997). Volume and number of neurons of the human hippocampal formation in normal aging and Alzheimer's disease. *The Journal of Comparative Neurology*, *494*(April 1996), 482–494.

Small, S. A., Schobel, S. A., Buxton, R. B., Witter, M. P., & Barnes, C. A. (2011). REVIEWS A pathophysiological framework of hippocampal dysfunction in ageing and disease. *Nature Reviews Neuroscience*, *12*, 585–601.

Verbeke, G. (1997). Linear mixed models for longitudinal data. In *Linear Mixed Models in Practice* (pp. 63–153). New York: Springer.

Vermetten, E., Vythilingam, M., Southwick, S. M., Charney, D. S., & Bremner, J. D. (2003). Long-Term Treatment with Paroxetine Increases Verbal Declarative Memory and Hippocampal Volume in Posttraumatic Stress Disorder. *Biological Psychiatry*, *54*, 693–702.

Videbech, P. & Ravnkilde, B. (1957). Reviews and overviews hippocampal volume and depression : a meta-analysis of MRI studies. *American Journal of Psychiatry*, *161*, 1957–1966.

Whelan, C. D., Hibar, D. P., Velzen, L. S., Van Zannas, A. S., Carrillo-Roa, T., Mcmahon, K., … Thompson, P. M. (2016). NeuroImage heritability and reliability of automatically segmented human hippocampal formation subregions. *128*, 125–137.

Wisse, L. E. M., Gerritsen, L., Zwanenburg, J. J. M., Kuijf, H. J., Luijten, P. R., Biessels, G. J., & Geerlings, M. I. (2012). NeuroImage Sub fields of the hippocampal formation at 7 T MRI : In vivo volumetric assessment. *NeuroImage*, *61*(4), 1043–1049.

Wisse, L. E. M., Jan, G., Heringa, S. M., & Kuijf, H. J. (2014). Neurobiology of aging hippocampal sub field volumes at 7T in early Alzheimer's disease and normal aging. *Neurobiology of Aging*, *35*(9), 2039–2045.

Yushkevich, P. A., Amaral, R. S. C., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., … Zeineh, M. M. (2015). NeuroImage quantitative comparison of 21 protocols for labeling hippocampal sub fields and parahippocampal subregions in in vivo MRI : towards a harmonized segmentation protocol. *Neuroimage*, *111*, 526–541.

Yushkevich, P. A., Wang, H., Pluta, J., Das, S. R., Craige, C., Avants, B. B., … Mueller, S. (2010). NeuroImage Nearly automatic segmentation of hippocampal sub fi elds in in vivo focal. *NeuroImage*, *53*(4), 1208–1224.

How to cite this article: Worker A, Dima D, Combes A, et al. Test–retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and Alzheimer's disease populations. *Hum Brain Mapp*. 2018;00:1–12. https://doi.org/10.1002/hbm.23948