Longitudinal Automatic Segmentation of Hippocampal Subfields (LASHiS) using Multi-Contrast MRI

Thomas Shaw[1*], Steffen Bollmann[1], Ashley York[2], Maryam Ziaei[1], Markus Barth[1, 3]

[1] Centre for Advanced Imaging, The University of Queensland, Brisbane, Australia
[2] Queensland Brain Institute, The University of Queensland, Brisbane, Australia
[3] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

*Corresponding Author
t.shaw@uq.edu.au
Address: Centre for Advanced Imaging, Building 57, Research Road, St Lucia, Brisbane, 4072, Australia

Acknowledgements

The authors acknowledge the facilities and scientific and technical assistance of the National Imaging Facility, a National Collaborative Research Infrastructure Strategy (NCRIS) capability, at the Centre for Advanced Imaging, The University of Queensland. MB acknowledges funding from Australian Research Council Future Fellowship grant FT140100865. This research was undertaken with the assistance of resources and services from the Queensland Cyber Infrastructure Foundation (QCIF). The authors would like to acknowledge Siemens Healthineers for the support of our project through the provision of the WIP sequence used to acquire the tse_UHF_WIP729C (variant: tse2d1_9) and MP2RAGE sequence (WIP 900) data in this publication.

Keywords: Hippocampus; longitudinal studies; segmentation; magnetic resonance imaging; image processing, computer-assisted

Abstract

The volumetric and morphometric examination of hippocampus formation subfields in a longitudinal manner using *in vivo* MRI could lead to more sensitive biomarkers for neuropsychiatric disorders and diseases including Alzheimer's disease, as the anatomical subregions have different roles. Longitudinal processing allows for increased sensitivity due to reduced confounds of inter-subject variability and higher effect-sensitivity than cross-sectional designs. We examined the performance of a new longitudinal pipeline (Longitudinal Automatic Segmentation of Hippocampus Subfields [LASHiS]) against three freely available, published approaches. LASHiS automatically segments hippocampus formation subfields by propagating labels from cross-sectionally labelled time-point scans using joint-label fusion to a non-linearly realigned 'single subject template', where image segmentation occurs free of bias to any individual time-point. Our pipeline measures tissue characteristics available in *in vivo* ultra-high resolution MRI scans and differs from previous longitudinal segmentation pipelines in that it leverages multi-contrast information in the segmentation process. LASHiS produces robust and reliable automatic multi-contrast segmentations of hippocampus formation subfields, as measured by higher volume similarity coefficients and Dice coefficients for test-retest reliability and robust longitudinal Bayesian Linear Mixed Effects results. All code for this project including the automatic pipeline is available at https://github.com/thomshaw92/LASHiS.

1. Introduction

The hippocampus formation is a brain structure generating large interest and research activity due to its implication in memory, psychiatric and neurological disorders including Alzheimer's Disease (AD; Daulatzai, 2013; Fotuhi, Do, & Jack, 2012), Motor Neurone Disease (Machts et al., 2018) and depression (Sapolsky, 2001), and especially its functional and structural changes in ageing (Fraser, Shaw, & Cherbuin, 2015). Due to the hippocampus formation's vulnerability in neurodegenerative disease, and its reported involvement in neurogenesis in the dentate gyrus (DG; Eriksson et al., 1998), precise volumetric and morphometric measurements of hippocampus formation are highly important for clinical studies and ageing research. Recent work has focussed on the hippocampus formation subfields or laminae, which are impacted differentially in neurodegeneration and disease (e.g., Machts et al., 2018). Volumetric and morphometric examination of these hippocampus subfields, especially in longitudinal studies, may lead to more sensitive biomarkers of disorder and the progress of the diseases (Boutet et al., 2014; Henry et al., 2011; Kerchner et al., 2012; La Joie et al., 2013; Maruszak & Thuret, 2014; Pluta, Yushkevich, Das, & Wolk, 2012).

Hippocampus subfields are functionally and cytoarchitectonically disparate (Andersen, 2007; Daulatzai, 2013; Fotuhi et al., 2012) with heterogeneous cellular composition. The four Cornu Ammonis (CA) subfields each have regional variations in pyramidal cells, creating structural differences between these subfields, which can be reflected to a degree in differing contrast and intensity signals in magnetic resonance imaging (MRI) scans with sufficient sensitivity and spatial resolution, i.e., at high enough field strengths (3 Tesla [T] and above; e.g., Duvernoy, Cattin, Risold, Vannson, & Gaudron,

2013; Naidich et al., 2003). As distinct cellular differences between subfields are only observable *ex vivo* and translate into only subtle differences in the MR signal, it is difficult to characterize these tissue classes at lower field strengths and routine MRI sequences due to low signal to noise ratio (SNR) and imaging artefacts (Wang & Doddrell, 2005).

Following from the above challenges, changes in small brain structures have been successfully studied using Ultra-High Field (UHF) MRI sequences with different contrasts (multi-contrast MRI) and allowed remarkable details for imaging *in vivo* (Fracasso et al., 2016). UHF MRI enables the increased spatial resolution necessary to characterize tissue differences *in vivo*, and in reasonable acquisition times. Previous UHF *in vivo* hippocampus subfield segmentation studies (for review, see Giuliano et al., 2017) utilise 'dedicated' sequences (e.g., single- or multi-echo Gradient Echo, Turbo-Spin Echo [TSE]) that exhibit different intensity and contrast characteristics for different tissue classes due to multiple refocusing pulses. Consequently, the laminae of the hippocampus are observable in these dedicated sequence types (Marques & Norris, 2018; Winterburn et al., 2013).

Advances in MRI acquisition techniques and image analysis methods have made automatic segmentation of hippocampus subfields possible. More recently, fully automatic hippocampus subfield pipelines including Freesurfer's hippocampus subfields method (Iglesias et al., 2015) and Automatic Segmentation of Hippocampus Subfields (ASHS; Yushkevich et al., 2015) have been developed using open-source segmentation software that combine several computational methods to achieve more reliable and precise results.

While both Freesurfer and ASHS have been applied in various studies (Andrea Chiappiniello, 2018; Iglesias et al., 2016; Pluta et al., 2012; Yushkevich et al., 2015), generally, segmentation errors cannot be avoided in practice. A study from Wisse et al. (2014) found that Freesurfer's method may be unreliable, with segmentations differing substantially from anatomical truths (e.g., CA2 and 3 consistently being reported as larger than CA1). The cross-sectional variant of Freesurfer accounts for contrast differences in input images while leveraging a combination of T1w and T2w contrasts for defining hippocampus segmentation. The underlying assumption of the Freesurfer scheme is that the spatial distribution of brain structures will be consistent with the *ex vivo* data in the atlas package, and spatial distributions of brain structures are homogenous within all scanned populations. A longitudinal variant of the hippocampus subfield method from Freesurfer (Iglesias et al., 2016) has also been introduced, which decreases residual (within-subjects) variability by allowing each participant to act as their own control. However, this method does not incorporate T2w information for labelling. It has been shown previously that T1w information generally does not contain signal that differentiates hippocampus subfields (Winterburn et al., 2013), including - in T2w contrast - the hypointense band of cells that separates the dentate gyrus (DG) from the CA regions known as the *stratum radiatum lacunosum moleculare*.

Longitudinal processing allows for increased sensitivity (Fitzmaurice, Laird, & Ware, 2011) due to reduced confounds of inter-subject variability and higher effect-sensitivity than cross-sectional designs. In image processing pipelines, longitudinal processing avoids many issues of secular trends inherent to cross-sectional designs, as participants act as their own control. These designs often exploit the knowledge that

within-subject anatomical changes over time are usually significantly smaller in scale than changes on an inter-subject morphological scale (Reuter, Schmansky, Rosas, & Fischl, 2012). Longitudinal designs have been used to successfully characterise changes in brain morphometry over time with greater accuracy than their cross-sectional counterparts (Reuter et al., 2012; Tustison et al., 2017). In particular, these designs avoid many types of image processing bias by transforming images into an intermediate space between time points where interpolation-related blurring is common across the time points.

Currently, using ASHS to measure volumes of hippocampus subfield in a single participant at multiple time-points does not account for the inherent variability present in cross-sectional methods. The Freesurfer longitudinal hippocampus subfields pipeline is the only dedicated longitudinal pipeline for measuring the volume of hippocampus subfield automatically. However, this method does not utilise the signal and tissue information available with multi-contrast MRI, and in particular, the 'dedicated' T2w scan commonly used in measuring the laminae of the hippocampus. We aimed to develop a longitudinal automatic hippocampus subfield segmentation pipeline that incorporates multi-contrast information while being robust to computational errors inherent to purely cross-sectional methods. We then examined the performance of our new longitudinal pipeline (Longitudinal Automatic Segmentation of Hippocampus Subfields [LASHiS]) against three published approaches viz; cross-sectional (FS Xs) and longitudinal (FS Long) Freesurfer hippocampal subfields (V6.0 Dev20181125; Iglesias et al., 2016), and ASHS cross-sectional (ASHS Xs; Yushkevich et al., 2015).

We developed an open-source multi-contrast pipeline that shares commonalities with existing pipelines, but is able to capture multi-contrast information from MRI scans automatically, while avoiding errors common to cross-sectional processing. We integrate a number of open-source software packages and programs to construct LASHiS, and propose the usage of multi-atlas fusion techniques to bootstrap automatic segmentation performance. Our pipeline is implemented with existing tools available through ANTs (ANTs Version: 2.2.0.dev116-gabc03; http://stnava.github.io/ANTs/; Avants, Tustison, & Song, 2010) and ASHS (https://sites.google.com/site/hipposubfields/; Yushkevich et al., 2015). Our pipeline and all associated code can be found at https://github.com/thomshaw92/LASHiS.

2. Methods and materials

2.1 Towards Optimising MRI tissue ChAracTerisation (TOMCAT) imaging data

Seven healthy participants (age: M = 26.29, SD = 3.35) were scanned using a 7T whole-body research scanner (Siemens Healthcare, Erlangen, Germany), with maximum gradient strength of 70 mT/m and a slew rate of 200 mT/m/s and a 7 T Tx/32 channel Rx head array (Nova Medical, Wilmington, MA, USA) in three sessions with three years between session one and two, and 45 minutes between two and three, allowing for a scan-rescan condition. Participants were scanned using a 2D TSE sequence (Siemens WIP tse_UHF_WIP729C, variant: tse2d1_9), TR: 10300ms, TE: 102ms, FA: 132°, FoV: 220mm, voxel size of 0.4 x 0.4 x 0.8mm^3 Turbo factor of 9; iPAT (GRAPPA) factor 2, acquisition time (TA) 4 minutes 12 seconds. The scan was repeated thrice over a slab aligned orthogonally to the hippocampus formation. An anatomical whole-brain T1w using a prototype MP2RAGE sequence (WIP 900; Marques et al., 2010; O'Brien et al., 2014) at 0.75mm isotropic voxel size was also

acquired (TR/TE/TIs = 4300ms / 2.5ms / 840ms, 2370 ms, TA = 6:54). At the first time-point, the nominal resolution was 0.5mm isotropic with the same parameters. For all subsequent processing, all MP2RAGE images for the first time-point were resampled to 0.75mm isotropic using b-spline interpolation. TSE images were resampled to 0.3mm isotropic and motion-corrected using non-linear realignment (Shaw et al., 2019) to ensure all segmentation strategies had the best and equivalent chance of succeeding.

2.2 Longitudinal Assessment of Hippocampal Subfields (LASHiS)

2.2.1 Atlas Construction

The entire LASHiS pipeline is described in Figure 1. Optionally, the ASHS pipeline can be optimised through the incorporation of a group-specific atlas. Similarly, creation of a group-specific atlas is a boon to our proposed method. This atlas is comprised of a representative pool of subjects (approximately 20-30 participants), manually labelled, and passed through the ASHS_train pipeline that is detailed in Yushkevich et al. (2015). Essentially, the manual segmentations are used as inputs (priors) for the joint-label fusion (JLF) algorithm in subsequent segmentations, and to train classifiers for the ASHS cross-sectional pipeline. Creating a group-specific atlas (of 20-30 subjects) would be beneficial for large longitudinal studies, as segmentation training would be performed on group-specific characteristics. However, having a group specific atlas is generally not necessary for robust performance of ASHS (Xie et al., 2018).

2.2.2 Preprocessing and cross sectional processing

The ASHS Xs pipeline has been previously proposed and discussed (Yushkevich et al., 2015). Briefly, the pipeline labels hippocampus subfields of a given T1w and

dedicated T2w scan covering the hippocampus subfields. This approach leverages a multi-atlas segmentation method and corrective learning techniques to segment (usually 3T or 7T) MRI data. The process involves first training existing manually labelled *in vivo* atlases of T2w scans. These trained atlas packages inform label for new *in vivo* T1w and dedicated T2w scans. ASHS provides many of these atlases at https://www.nitrc.org/projects/ashs. These publically available atlases may be replaced with a group-specific atlas as in 2.1.1. The T2w input scan is usually acquired anisotropically with reduced resolution along the major axis of the hippocampus subfield and high in-plane resolution. The spiral structure of the hippocampus formation does not change rapidly along its major axis, which motivates this parameter choice (Iglesias et al., 2016). ASHS Xs employs similarity-weighted voting for learning segmentation priors and JLF for multi-atlas segmentation prior classification. In the segmentation protocol, weighted voting at the voxel level derives 'strong' segmentation choices for the target image (Yushkevich et al., 2015).

For preprocessing of all data, we included modified preprocessing steps based on the ANTs cortical thickness pipeline (Tustison et al., 2014) and our previous work (Shaw et al., 2019). These steps were incorporated to ensure consistent segmentation results across participants and included:

I. Skull stripping (i.e., ROBEX; Iglesias, Liu, Thompson, & Tu, 2011) of the T1w scan for removal of background tissue and artefacts that may result in registration errors further downstream
II. N4 bias correction (ANTS version 2.20.dev116-gabc03; Tustison et al., 2010) of the T1w scan that mitigates low spatial frequency variations in the data

III. Rician denoising of T1w and T2w scans (Manjón, Coupé, Martí-Bonmatí, Collins, & Robles, 2010), which has been shown to reduce high-frequency Rician noise in MRI scans (Tustison et al., 2017)

IV. Intensity normalisation of T1w and T2w scans. We utilized 'NiftiNorm' (https://github.com/thomshaw92/nifti_normalise), a nifti implementation of 'mincnorm' from the medical imaging network common (MINC) data toolkit (Vincent et al., 2016) that normalises image intensities between two percent-critical thresholds, removing outlying intensity values

V. If multiple repetitions of the dedicated T2w scan are available, non-linear realignment of these scans to reduce motion artefacts and increase the sharpness of the scans as in Shaw et al. (2019).

LASHiS derives initial segmentations of each time point cross-sectionally using the ASHS pipeline with an atlas package similar to the subject pool's intensity and spatial characteristics. This yields an atlas-defined number of hippocampus subfield labels. Due to our small sample size, it was not possible to create a bespoke atlas for validation. Therefore, we utilized ASHS (V2.0) with the Penn Memory Center 3T ASHS Atlas (Yushkevich et al., 2015). Despite the difference in field strength between our data and the atlas data, Xie et al (2018) has found that atlas composition does not significantly affect segmentation between 7T and 3T, and the contrast and intensity profiles of the scans in the 3T atlas are similar to the TSE scans we collected in the present study.

2.2.3 Single-subject template (SST)

In parallel to cross-sectional processing, a minimum deformation average (MDA) multi-contrast template of average intensity and shape is created in accordance with Avants et al. (2010). This template serves as an intermediate between any *n* time points of a subject and is biased equally to any given time-point. All subsequent processing of hippocampus volumes is done in the space of the SST in order to treat all time points in the same way. We have also found previously that combining scans in this way increases segmentation consistency and image sharpness (Shaw et al., 2019).

2.2.4. Joint-Label Fusion (JLF) and longitudinal estimations of segmentations

Following SST creation and labelling, and cross-sectional labelling, individual time-point multi-contrast scans and their cross-sectionally defined segmentation labels then act as multi-contrast atlases to compute SST labels using JLF. The intended usage of JLF is to propagate manually derived labels to a target image. However, we used JLF with the atlases being *automatically* labelled. We also include the automatically labelled SST as an extra input to increase the power of the method. JLF assigns the spatially varying atlas (input) weights to the SST in a way that accounts for error correlations (Wu et al., 2017) between every *n* pairs of atlases. In this way, no single time point is biased towards the segmentation of the SST, and the SST is labelled based on a weighted vote of the segmentations from each time-point and the SST. In our scheme, a working region of interest (ROI) is defined roughly around the hippocampus, non-linearly warped to the space of the SST ROI, and JLF applied with parameters chosen based on Wang et al., (2013). The inverse of these non-linear transformations is later used for labelling the input images. This approach, therefore, bootstraps cross-sectional segmentations of hippocampus subfield to the SST, and

the best fitting labels are chosen based on the intensity and shape characteristics of the SST, not the individual time-points.

Subsequently, the inverse of each of the time-point to SST transformations calculated by the registration from time-point to SST in the JLF piecewise registration is applied to the newly generated SST hippocampus subfield labels individually, warping the SST labels to each individual time-point. Provided the time-point-to-SST registrations are accurate (Avants et al., 2011) and invertible, the reverse normalisation of the labels can be considered a robust and reliable method for transforming the labels to the space of the subject's time-point hippocampus subfield labels. Finally, we used Convert3D (Yushkevich et al., 2006) to measure the new subfield volumes in time-point space.

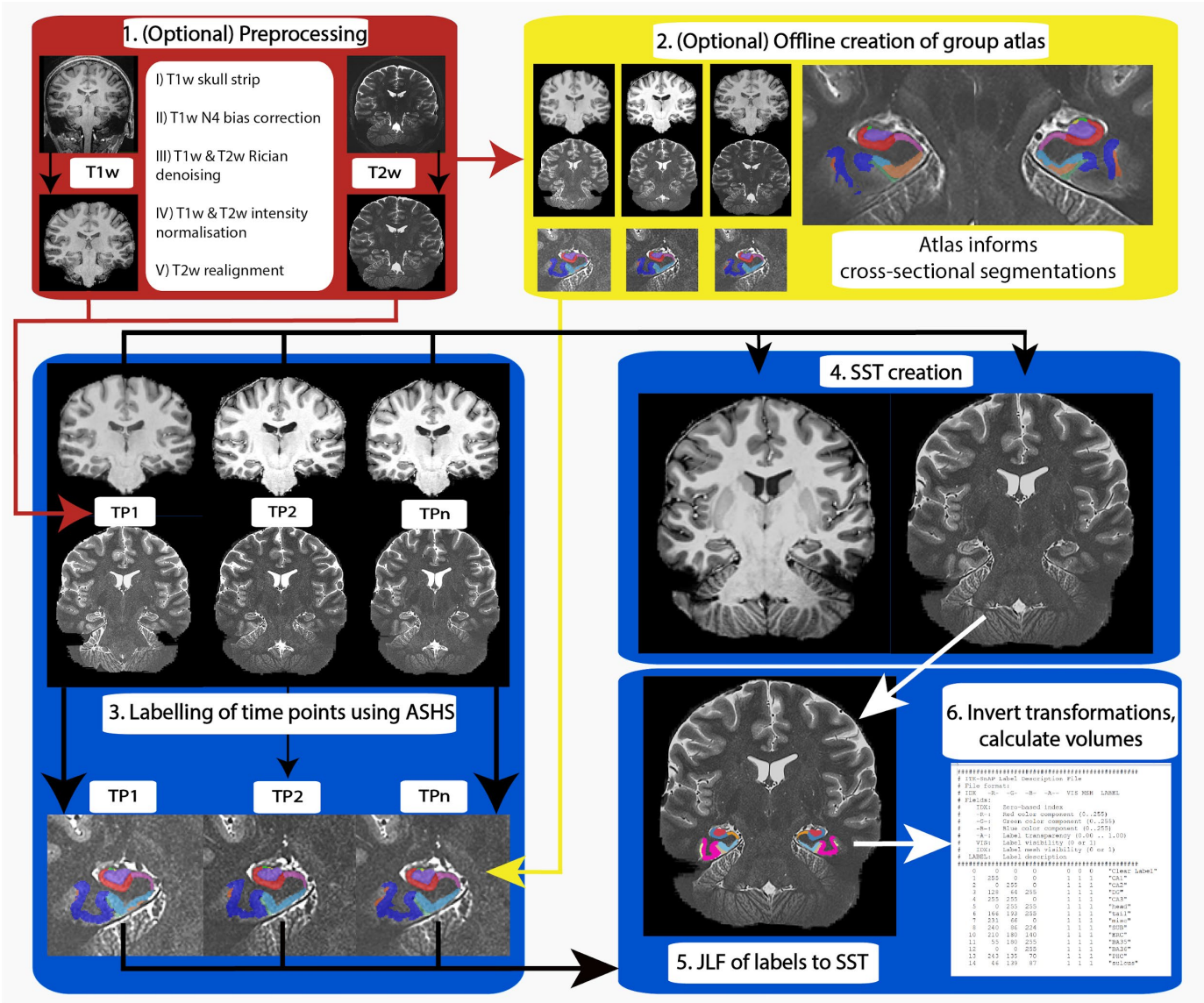


Figure 1. Longitudinal Automatic Segmentation of Hippocampus Subfields (LASHiS) pipeline. The pipeline consists of the following steps with the input of any number of T1w and T2w individual time points per participant: 1) (In red) (optional) preprocessing of both T1w and T2w scans. 2) (In yellow) Offline construction of a sample-specific atlas for LASHiS. 3) Labelling of individual time points of each subject using ASHS and a representative atlas (or an atlas created in [2]) to yield hippocampus subfield estimates. 4) Construction of a multi-contrast Single Subject Template (SST). 4) JLF of each of the individual time point labels to the SST using both contrasts and individual hippocampus subfield labels to produce a labelled SST. 5) Application of the inverse subject-to-SST transformations to SST labels. 6) Measurement and calculation of subfield labels in subject-space.

2.3 Statistical Evaluation

2.3.1. Hippocampus subfield segmentation methods

We compared the performance of our LASHiS strategy (as detailed above) with three other established strategies, and one other exploratory strategy detailed below, examples of the output of each segmentation strategy are given in Figure 2:

1. Cross-sectional ASHS (ASHS Xs) The segmentation strategy described in Yushkevich et al. (2015) used to compute segmentations for each timepoint independently in a cross-sectional manner for each time-point. We utilised segmentation results that incorporated the high-resolution T2w scan information. We modified certain parameters in ASHS to account for our 7T high-resolution and preprocessed (isotropic) inputs to account for resolution and image size. We used the Penn Memory Centre atlas https://www.nitrc.org/frs/?group_id=370 (Yushkevich et al., 2015) for segmentation due to input-atlas contrast similarities and an increased number of subfield label outputs compared to the available 7T atlases.
2. Cross-sectional Freesurfer hippocampus subfield segmentation (FS Xs): the method described in Iglesias et al., (2015) was used to compute segmentations for each timepoint independently in a cross-sectional manner. Due to skull strip failures in recon-all and mri_watershed, the brain mask was replaced with the brain mask created in the preprocessing steps using ROBEX (Iglesias et al., 2011) in order to give Freesurfer the best chance of succeeding.
3. Freesurfer longitudinal hippocampal subfields (FS Long): This pipeline, described by Iglesias et al. (2016) utilises intensity and contrast information from an *ex vivo* manually traced atlas of hippocampal subfields to delineate *in vivo* subfield information. The *ex vivo* atlas is supplemented by an *in vivo*

Freesurfer atlas (as described in Kennedy, Filipek, & Caviness, 1989), which informs segmentation of the geometric priors surrounding the hippocampus. In this way, the generative model that classifies hippocampal subfields *in vivo* is calculated from the spatial distribution of the hippocampus and its surrounding brain regions as described in the atlas priors.

4. JLF-free LASHiS (Diet LASHiS): This method is similar to LASHiS, though does not utilise the JLF bootstrapping step or cross sectional processing, thus reducing processing time by approximately 20%. Instead, the SST is created, labelled using ASHS and a representative atlas package, and labels were reverse-normalised to time-point space using the transformations calculated in the template building procedure, as distinct from the subject-SST transformations derived in the JLF step. We incorporated this method to determine the relative importance of the JLF bootstrapping step in our pipeline. This was considered a standard reverse normalisation segmentation pipeline.

2.3.2 Evaluation methods

Here, we evaluate our strategy in line with previously published methods in order to quantify reliability, reproducibility and precision. We reproduced analyses employed by both the FS Long hippocampus segmentation strategy (i.e., test-retest reliability) and longitudinal Bayesian Linear Mixed Effects (LME) modelling (Sorensen, Hohenstein, & Vasishth, 2016).

2.3.3 Experiment one: Test-retest reliability

We evaluated the test-retest reliability of all methods through testing differences between the second and third acquired time-point scans in the TOMCAT dataset. For

each participant, we segmented each scan-rescan session with the five segmentation methods and assessed performance based on two metrics: 1) absolute differences in volume estimates for each hippocampus subfield label between scan-rescan acquisitions, and 2) Dice overlap (Dice, 1945). We first measured the volume similarity coefficient, which does not rely on segmentation locations (Taha & Hanbury, 2015). This metric does not implicitly rely on overlaps in segmentations (such as Dice overlaps, which can be difficult to measure without bias when comparing between analysis strategies, as in Iglesias et al., [2016]). For completeness, and to have a direct comparison with Iglesias et al. (2016), we also assessed Dice overlaps between time-point two and three in all pipelines. The Dice coefficient between two binary masks is defined in its simplest form as "twice the number of elements common to both segmentations divided by the sum of the elements in the segmentations", and is described as:

$$1.\quad DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

where a perfect overlap between two segmentations (*X* and *Y*) is 1, and no overlap is 0 (Taha & Hanbury, 2015). In LASHiS, Diet LASHiS, FS Xs, and ASHS Xs, the final result of hippocampus subfield labels occurs in a native (input) space. We linearly resampled all labels in these four pipelines to an intermediate space (SST space), and calculated Dice overlaps in these cases with the fuzzy Dice counterpart using the freely available EvaluateSegmentation tool (Taha & Hanbury, 2015). There is a bias towards FS Long for having superior Dice overlap evaluation due to the extra interpolation required in these linear realignments, which are not necessary in FS Long. We discuss the implications of this in section 4.1.

We leveraged Bayesian paired *t*-tests in accordance with Rouder, Speckman, Sun, Morey, and Iverson (2009) to assess the differences in subfield changes across the second and third time-point using Jamovi, R, and the BayesFactor plugin (Morey & Rouder, 2019; R Core Team, 2019; The Jamovi Project, 2019). In our analyses, $BF_{10} > 3$ was taken as substantial evidence for the alternative hypothesis, with $BF_{10} > 10$ taken as strong evidence, and BF_{10} greater than 100 were considered decisive. BF_{10} values between 1 and 3 were considered anecdotal evidence for the alternative hypothesis. In contrast, $BF_{10} < 0.33$ (or $BF_{01} > 3$) was considered as substantial evidence for the null, with BF_{10} between 0.33 and 1 providing anecdotal evidence for the null hypothesis in accordance with Lee & Wagenmakers (2013).

2.3.4 Experiment two: Bayesian Longitudinal Linear Mixed Effects Modelling

To assess relationships between cross-sectional and longitudinal results while accounting for subject-specific trends (Tustison et al., 2017), we quantify between, and within (residual) variability of hippocampus subfield volume. In this experiment, we aimed to assess the utility of each pipeline for measuring each hippocampus subfield and detecting potential biomarkers therein. It is possible to quantify the relative performance of cross sectional and longitudinal pipeline variants with bayesian LME models (Tustison et al., 2017). Intuitively, the best longitudinal method maximises both within-subject reproducibility and between-subjects variability (to distinguish between sub-groups). Good performance is quantified by maximising the ratio between between-subject variability and residual variability. A summary measure of this is the variance ratio, which shows the linear relationship between within- and between-subjects variability, which is a useful measure of performance for longitudinal

pipelines. Higher variance ratios indicate optimised prediction and confidence intervals for the segmentation quality.

Freesurfer and ASHS provide different outputs in terms of subfield names. To overcome difficulties computing variance values relating to non-overlapping regions, we have concatenated several subfields that share commonalities across all pipelines and present these in Table 1. Subfields that did not share any commonalities across pipelines were excluded from analysis.

Table 1. *Label names for all hippocampus subfields that share similarities between Freesurfer and ASHS pipelines.*

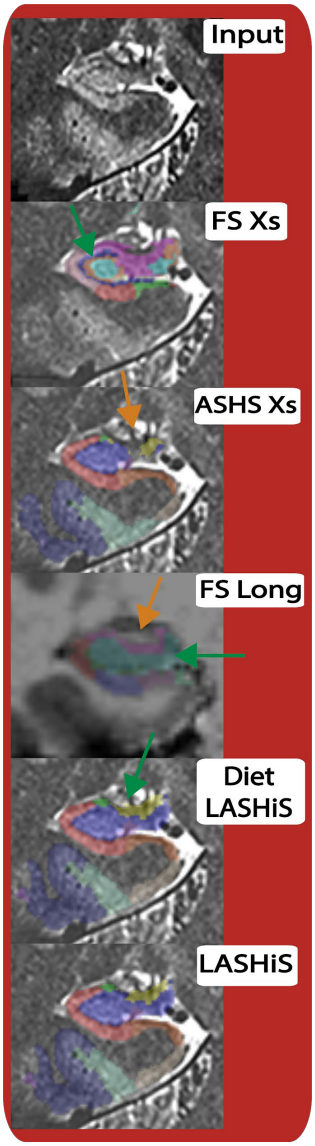
Combined label name for analysis	Freesurfer label names (FS Xs, FS Long)	ASHS label names (ASHS Xs, Diet LASHiS, LASHiS)
CA1	CA1-head & CA1-body	CA1
CA2-3	CA3-head & CA3-body†	CA2 & CA3
DG	GC-ML-DG-head, GC-ML-DG-body, CA4 head, & CA4 body	DG‡
SUB	Presubiculum-head, presubiculum-tail, subiculum head, & subiculum tail	SUB

†Freesurfer combines estimates of CA2&CA3 as label CA3 in their algorithm.

‡ASHS combines estimates of DG and CA4 as label DG in their algorithm.

The subfields chosen in our analysis include Cornu Ammonis (CA) region 1 (CA1), CA region 2 and 3 (which was combined in Freesurfer's method; CA2-3), Subiculum (SUB; comprising presubiculum and subiculum in the Freesurfer pipeline), and dentate gyrus (DG; comprising of CA4 and DG in the ASHS atlas package). These four subfields are measured for all analyses for left and right sides for a total of eight subfields. Note that LASHiS computes as many labels as are included in the initial atlas package (usually 14 per side). For consistency of subfield volumes that may be influenced by intracranial volume (ICV), and to obtain more internally consistent measures of volume, we normalised all raw volume values by total hippocampus formation volume (e.g., CA + DG + SUB). We examined subfield results for all comparisons, but report significant differences only between the most relevant comparisons: namely between LASHiS and FS Long, as these are the two longitudinal pipelines of interest.

Figure 2. Hippocampus subfield segmentation results (coloured) for a single representative subject for the five tested methods at the same slice in a coronal view. Each segmentation result is overlaid on the high-resolution T2w scan save for FS Long, which utilises a T1w scan for segmentation. Green arrows denote a possible under-segmentation, orange for a possible over-segmentation of a subfield where subfield information is not available.



3.0 Results

3.1 Experiment one: test-retest reliability

We conducted a series of Bayesian paired-sample t-tests in order to test absolute volume differences between the second and third time-point. Figure 3 shows differences between methods for volume similarity in the test-retest experiment. We found that LASHiS and Diet LASHiS showed significantly higher volume similarity in all subfields than other methods. Specifically, We found LASHiS to have *decisively* higher ($BF_{10} > 100$) volume similarity coefficients compared to FS Long in all subfields.

ASHS Xs also showed high volume similarity in DG subfields compared to the Freesurfer variants, though with high variability; we observed larger variability in the volume similarity in all other methods compared to LASHiS variants (see Supplementary Figure 1 for subfield variance breakdowns). All other comparisons with LASHiS are detailed in Supplementary Figure 1 and 2.

We next conducted Bayesian paired-sample *t*-tests for Dice overlaps between the segmentation labels in the second and third time-point. Figure 3 shows Dice overlap values of each subfield for each method. Note, that Dice scores for LASHiS, Diet LASHiS, Freesurfer Xs and ASHS Xs are negatively affected by the resampling needed to compute the registrations between the two time-points, which is not present in the FS Long method. Interestingly, our results do not directly replicate Iglesias et al. (2016) in terms of mean Dice overlap scores for test-retest reliability. We found slightly lower Dice overlaps for all subfields in our sample in the Freesurfer methods compared to Iglesias et al. (2016).

In terms of subfield differences, we will detail comparisons only for LASHiS and FS Long, with all method comparisons included in Supplementary Figure 3 and 4. We found that Dice overlaps for LASHiS were higher than FS Long for test-retest reliability *decisively* in Left-DG and Right-DG ($BF_{10} > 100$) and *anecdotally* in Left-CA1 ($BF_{10} > 1$). We found no difference between LASHiS and FS Long in Right-CA1, and Right-SUB ($BF_{10} < 1$). FS Long had *substantially* higher scores than LASHiS for Right-CA2-3, Left-CA2-3, and Left-SUB ($BF_{10} > 10$).

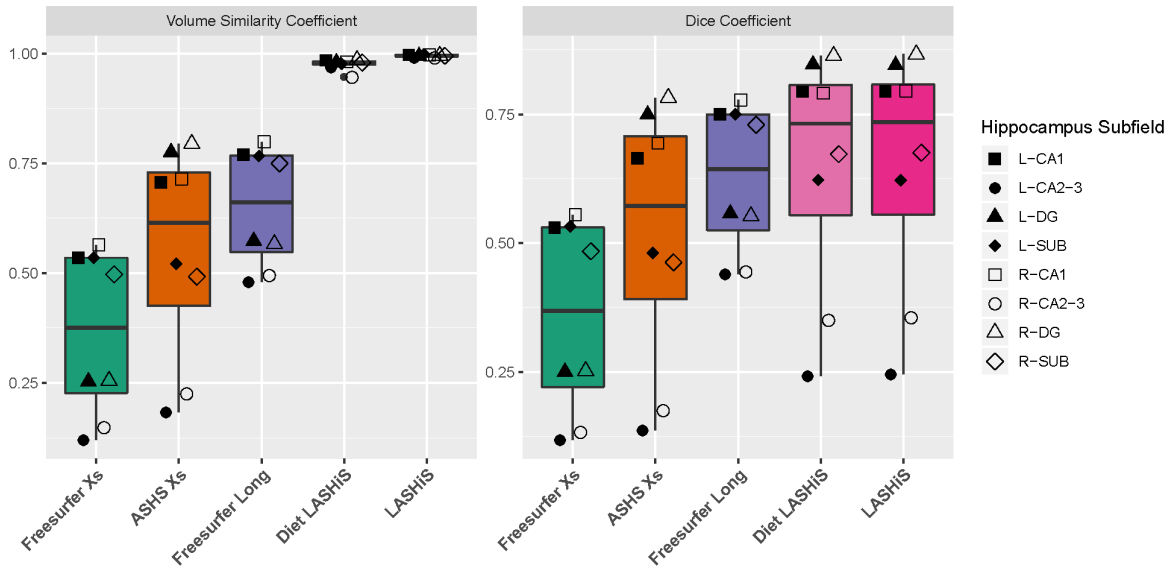


Figure 3. *Box plots of volume similarity coefficients (left) and Dice coefficients (right) of each hippocampus subfield (left, black filled shapes, and right, white filled shapes) from time-point two and time-point three for each method, where a value of 1 represents perfect overlap between time-points, and 0 represents no overlap. Error bars represent overall standard deviation. Freesurfer Xs, ASHS Xs, Diet LASHiS and LASHiS all require resampling to a common space before overlap calculation of Dice overlaps. Higher scores between time-points denote higher subfield overlap between test-retest time-points.*

3.2 Experiment Two: Bayesian longitudinal Linear Mixed Effects modelling

We compared the performance of five hippocampus subfield segmentation processing approaches using longitudinal LME modelling to quantify between and residual variability, and the variance ratio of these. Figure 4 provides the 95% confidence intervals for the variance ratios in each subfield for each of the pipelines. As noted in Tustison et al. (2017), “superior methodologies are designated by larger variance ratios”. Across subfields, LASHiS has higher variance ratios for Left-CA1, Left-SUB, Right-DG, and Right-SUB. FS Long out-performs LASHiS for Left-CA2-3 and Right-

CA1 and ASHS Xs performs best for Right-CA2-3 and Left-DG (followed closely by LASHiS and diet LASHiS). We also note lower values in CA2-3 subfield variance ratios in LASHiS in both hemispheres.

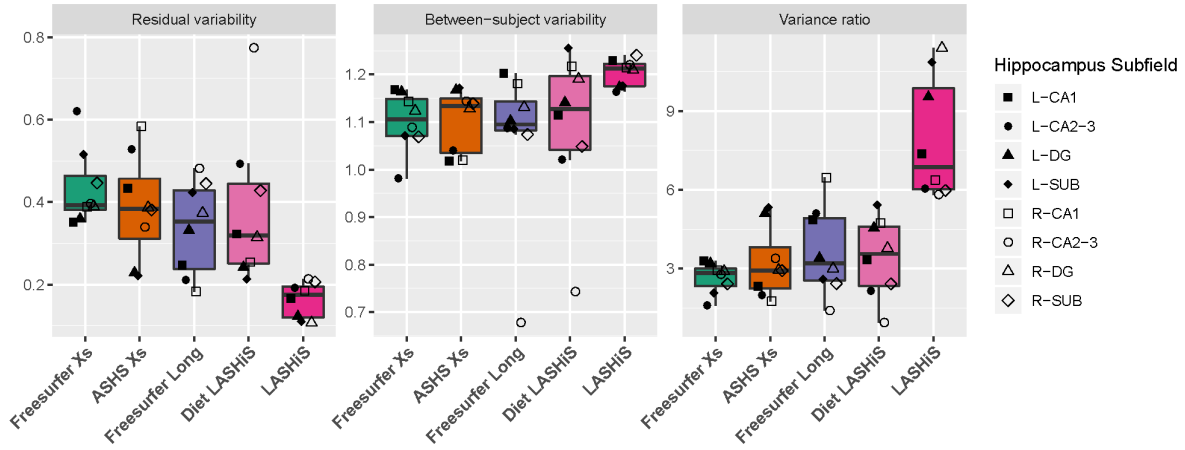


Figure 4. Box plots of (from left to right) residual variability, between-subject variability, and the variance ratio between the two variabilities of each method for the three time-points. Error bars denote the 95% confidence intervals. Shapes with black and white fill represent individual left, and right hippocampus subfields, respectively. Lower residual, and higher between-subject variabilities are preferred for longitudinal pipelines. The variance ratio is a summary statistic of the two variability values, with higher values indicating improved discrimination between within- and between-subject variance.

We found overlapping confidence intervals for all pipelines, with obvious trends towards LASHiS as having the highest variance ratios. Figure 4 (right) shows the relative distributions of variabilities (between and within variability, and variance ratio) collapsed across subfields for each of the assessed pipelines. A trend towards higher variance ratio for LASHiS compared to the other methods can be observed, and that

the contributions of the residual and between subject variability accounts for the variance ratios observed. We also note that variance ratios for Diet LASHiS are comparable to FS Long. Figure 4 shows variance ratios of each pipeline broken down by subfield. There are clear trends towards LASHiS having the superior variance ratios in Left and Right CA1, Left and Right SUB, and Right DG regions. LASHiS has low variance ratio values in Right CA2/3. LASHiS therefore has higher variance ratios than all other methods in 5 of 8 subfields. We also provide within and between subject variability for each subfield in Supplementary Figure 5 and 6, respectively, and the subfield variance breakdowns for the variance ratio in Supplementary Figure 7.

4.0 Discussion

4.1 Experiment One, test-retest reliability

The test-retest results highlight the reliability of the LASHiS pipeline. Capitalising on the availability of data from multiple time-points to increase SNR in the SST improves the inherent regularisation and prior information for segmentation, as proposed for LASHiS. LASHiS and Diet LASHiS show excellent test-retest reliability for volume similarity coefficients. Deformable registration has been previously used successfully to segment hippocampus structures in groups of participants (Hammers et al., 2007; Hogan et al., 2000). LASHiS benefits from deformable registration-based image segmentation, as the hippocampus is segmented only in SST space. We contrast LASHiS with FS Long, which utilises an SST in order to compute time-point segmentations. However, the FS Long SST uses only T1w information, potentially limiting the reliability of the time-point to SST registration, and therefore decreasing the volume similarity. Indeed, Iglesias et al. (2016) found an average of 4.5% difference in absolute volume similarity in their test-retest condition. In terms of Dice,

all other methods were at a disadvantage to FS Long for this metric due to the interpolation step required to realign the scans. This effect was mitigated through the utilisation of a fuzzy Dice coefficient in all other methods. However, despite the disadvantage, LASHiS shows comparable Dice overlaps in the test-retest condition to FS Long, except for in the smaller subfields (e.g., CA2/3). Coupled with the results of volume similarity, we can assert that LASHiS is a reliable method for longitudinal hippocampus subfield segmentation.

4.2 Experiment Two, Bayesian longitudinal Linear Mixed Effects modelling

Many evaluation strategies employ manual segmentations (e.g., Berron et al., 2017) to provide a gold standard for evaluation of any segmentation strategy. However, manual hippocampus subfield segmentation is time and labour intensive, taking up to eight hours initially, and two hours after five months of training (Wisse et al., 2016) and is prone to inter- and intra-rater variability (Boccardi et al., 2011; Hsu et al., 2002; Mulder et al., 2014). We explored the usefulness of LASHiS in the examination of the variance ratio in our longitudinal Bayesian LME modelling experiment. Higher variance ratios that are characterised by both lower residual variability and larger between-subjects variability are beneficial for longitudinal cohort studies. We found the highest variance ratios in LASHiS, underscoring the usefulness of our approach in maximising between subject differences. We note outliers in variance ratios in LASHiS and FS Long, which are driven by results in CA2-3, and SUB, respectively. For LASHiS, high residual variability was found for right CA2-3, driving this outlier.

We want to note here, that LASHiS is potentially negatively biased by limitations in subfield selection. All Freesurfer schemes combine CA2 & CA3 estimates in their

algorithms. In calculating our subfield estimates, we summed CA2 and CA3 volumes offline, potentially biasing our estimates of residual variability. We note a similar residual variability outlier in the ASHS Xs scheme in the left CA2-3 combined subfield regions. Volume estimates of CA2 and CA3 regions were generally reported less precisely than other subfields, as measured by the low test-retest statistics and low within-subject variability in the LME experiment. Previous research (Dalton, Zeidman, Barry, Williams, & Maguire, 2017; Pipitone et al., 2014; Wisse et al., 2016; Yushkevich et al., 2015) has repeatedly shown discrepancies in reporting these subfield boundaries *in vivo*. This is largely due to their small size and the reliance on heuristic geometric rules for segmenting CA2/3 subfields on *in vivo* MRI, rather than visible contrast differences in the scan. Thus, inter- and intra-rater reliability are often low for these subfields (Xie et al., 2018). Our automatically derived subfield estimates are likely influenced by discrepancies in the manual labels that inform segmentations. Notably, FS Long also suffers from a low variance ratio in CA2-3, suggesting either i) a homogeneous participant pool leading to low between-subject variability, or ii) large, unexpected differences in time-points in these subfields, or iii) a combination of these.

4.4 Benefits and advantages of LASHiS

Both LASHiS and the FS Long scheme segment hippocampus subfield and derive volume estimates from MRI images. However, only a T1w scan of an individual is processed through the longitudinal stream of ‘recon-all’ before longitudinal processing of hippocampus subfields, potentially explaining the FS Long results compared to LASHiS. Our design utilises multi-contrast information from MRI scans and importantly, our design allows for information that can *only* be captured by multi-

contrast MRI (i.e., the laminae of the hippocampus subfield) to be included in the labelling.

LASHiS derives its power from its ability to decrease random errors in the labelling procedure, and through increasing the likelihood for correct labelling to occur when the SST is created, which, due to the non-linear realignment between the time-points, implicitly increases SNR and sharpness of the SST compared to the individual time points through the template building procedure (Shaw et al., 2019). Our inclusion of Diet LASHiS highlights the contributions of the JLF step from the simple labelling of the SST, which may be subject to both random and systematic errors. In LASHiS, these random errors may be mitigated in part due to the bootstrapping of JLF from the individual time-point to the SST. It is possible this step decreases the likelihood of random errors in the labelling scheme because of JLFs ability to vote on labels that fit best to the SST. Therefore, random variance caused by mislabelling at any individual time-point may be ameliorated by the JLF step. In turn, this is the likely cause for the low residual variability found in LASHiS in Experiment Two. Our inclusion of JLF labelling using *automatically* generated labels is a novel consideration in the field of hippocampus subfield segmentation, and relies on the assumption that automatically generated subfield labels are considered accurate.

We included a computationally less expensive and faster approach to multi-contrast hippocampus subfield segmentation, namely Diet LASHiS. This method performs all steps save for the initial cross-sectional segmentations and the bootstrapping of these segmentations to the SST using JLF. Diet LASHiS performed well in the volume similarity portion of Experiment 1, and in Experiment 2 in comparison to the other

methods examined, though to a lesser degree than LASHiS. As the steps taken to complete LASHiS and Diet LASHiS are the same save for the additional JLF bootstrapping method in the former, we propose the increased sensitivity and robustness in the LASHiS scheme was due to the JLF step. Indeed, despite the disadvantage of potentially increasing systematic errors with the JLF bootstrapping step, it is evident that these systematic errors are largely overcome in the initial cross-sectional labelling of the hippocampus subfield with ASHS Xs.

Processing time for LASHiS depends largely on compute infrastructure, T2w image size, and the number of time-points. Our testing on three time-points with large ($0.3mm^3$) T2w images ran in the order of 24 hours on a single CPU core, and 6 hours on 12 cores without parallelisation. Many steps, including the initial cross-sectional segmentations and SST creation, can be run in parallel using job scheduling software (PBS, Sun Grid Engine, Slurm, etc.) and parallelised across cores, decreasing the time required in orders of magnitude less commensurate with the number of cores employed. Diet LASHiS is estimated to decrease compute time by approximately 20%, as neither the cross-sectional, nor the JLF steps are required. ASHS Xs takes between 1-2 hours on a single core, while FS Xs takes approximately 40 minutes after 48 hours of preprocessing using a single core (https://surfer.nmr.mgh.harvard.edu/fswiki/ReconAllRunTimes). Fs Long takes approximately 60 minutes on a single core after 48 hours of cross-sectional processing per time-point, and further creation of an SST. The great advantage of LASHiS is the flexibility of computational processing options for each step, allowing for scalable processing for larger datasets.

Our incorporation of a Bayesian approach to the widely used longitudinal LME method for examining differences in method performance aids in discrimination of subtle differences between participants with small variability (as in the present study). This technique allowed us to simultaneously examine small differences between participants, while also capturing longitudinal within-subject changes; both of which are especially important in examination of change in clinical subpopulations and other low-*n* studies, where small longitudinal changes need to be captured precisely. Notwithstanding the robustness of our analysis technique, larger studies with more variable neuroanatomy are required to show the true robustness of this pipeline.

4.5 Limitations

The design of our pipeline decreases random variability in any session due to the SST registration and JLF labelling scheme. A limitation of our scheme is that label errors (i.e., systematic errors) in subjects will propagate to the SST, despite the sophisticated JLF algorithm employed that does not independently compute similarity weights between the pairs while voting (Wang et al., 2013). Therefore, it is important to note that LASHiS is never free from labelling errors that occur in all image segmentation pipelines. These systematic errors can be avoided through quality assurance of scans and labels at the cross-sectional level (i.e., before the JLF bootstrapping step), which is essential in any volumetric labelling scheme, regardless.

We here report a small healthy cohort of young adults with no known psychological or neurological disorders. We assumed no difference between time-point two and three, and very small differences between time-point one and two due to the age and health of the participants. We concede this limitation in our interpretation of test-retest

analyses and further work is needed to examine LASHiS' ability to detect differences in larger cohorts of clinical and healthy subjects. The Bayesian nature of our longitudinal LME modelling accounts for small sample sizes (Sorensen et al., 2016), and the results of Experiment One and Two should therefore not be affected by our small sample size.

Our test-retest statistics show that LASHiS has improved metrics compared to other longitudinal methods, with obvious differences to previous work reporting the same methods (Iglesias et al., 2015), where Dice overlaps were considerably higher overall for the Freesurfer methods. We note this limitation of having such a small sample size in the present study, which was the likely reason for the higher variability in the Dice overlap scores in the Freesurfer method. However, as LASHiS shows a consistent improvement compared to all other methods, we are confident LASHiS is a robust and reliable method for longitudinal multi-contrast hippocampus subfield segmentation.

4.6 Conclusions

Here, we present a technique for automatically and robustly segmenting hippocampus subfield volumes using UHF multi-contrast MRI in healthy subjects. We found that LASHiS shows marked improvements across a number of relevant measures, such as Dice similarity and volume similarity coefficients for test-retest reliability, and Bayesian LME modelling, compared to other methods used for cross-sectional and longitudinal hippocampus segmentation. This is likely due to its utilisation of multi-contrast information that better captures hippocampus subfield tissue characteristics and its ability to decrease random errors in the labelling procedure.

References

Andersen, P. (Ed.). (2007). *The hippocampus book*. Oxford ; New York: Oxford University Press.

Andrea Chiappiniello. (2018, June 25). Multicentric test-retest reproducibility of human hippocampal volumes: FreeSurfer 6.0 longitudinal stream applied to 3D T1, 3D FLAIR and high-resolution 2D T2 structural neuroimaging. Retrieved June 26, 2018, from http://indexsmart.mirasmart.com/ISMRM2018/PDFfiles/3247.html

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, *54*(3), 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025

Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., & Gee, J. C. (2010). The Optimal Template Effect in Hippocampus Studies of Diseased Populations. *NeuroImage*, *49*(3), 2457. https://doi.org/10.1016/j.neuroimage.2009.09.062

Avants, B., Tustison, N., & Song, G. (2010). *Advanced Normalization Tools (ANTS)*. 35.

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

Boutet, C., Chupin, M., Lehéricy, S., Marrakchi-Kacem, L., Epelbaum, S., Poupon, C., … Colliot, O. (2014). Detection of volume loss in hippocampal layers in Alzheimer's disease using 7 T MRI: A feasibility study. *NeuroImage : Clinical*, *5*, 341–348. https://doi.org/10.1016/j.nicl.2014.07.011

Dalton, M. A., Zeidman, P., Barry, D. N., Williams, E., & Maguire, E. A. (2017). Segmenting subregions of the human hippocampus on structural magnetic resonance image scans: An illustrated tutorial. *Brain and Neuroscience Advances*, *1*,

2398212817701448. https://doi.org/10.1177/2398212817701448

Daulatzai, M. A. (2013). Neurotoxic Saboteurs: Straws that Break the Hippo's (Hippocampus) Back Drive Cognitive Impairment and Alzheimer's Disease. *Neurotoxicity Research*, *24*(3), 407–459. https://doi.org/10.1007/s12640-013-9407-2

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. https://doi.org/10.2307/1932409

Duvernoy, H. M., Cattin, F., Risold, P.-Y., Vannson, J. L., & Gaudron, M. (2013). *The human hippocampus: Functional anatomy, vascularization and serial sections with MRI* (Fourth edition). Heidelberg ; New York: Springer.

Eriksson, P. S., Perfilieva, E., Björk-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. https://doi.org/10.1038/3305

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis*. Retrieved from http://ebookcentral.proquest.com/lib/uql/detail.action?docID=1051443

Fotuhi, M., Do, D., & Jack, C. (2012). Modifiable factors that alter the size of the hippocampus with ageing. *Nature Reviews Neurology*, *8*(4), 189–202. https://doi.org/10.1038/nrneurol.2012.27

Fraser, M. A., Shaw, M. E., & Cherbuin, N. (2015). A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. *NeuroImage*, *112*, 364–374. https://doi.org/10.1016/j.neuroimage.2015.03.035

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), 481–494. https://doi.org/10.1002/hipo.22717

Hammers, A., Heckemann, R., Koepp, M. J., Duncan, J. S., Hajnal, J. V., Rueckert, D., & Aljabar, P. (2007). Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: A proof-of-principle study. *NeuroImage*, *36*(1), 38–47. https://doi.org/10.1016/j.neuroimage.2007.02.031

Henry, T. R., Chupin, M., Lehéricy, S., Strupp, J. P., Sikora, M. A., Sha, Z. Y., … Van de Moortele, P.-F. (2011). Hippocampal Sclerosis in Temporal Lobe Epilepsy: Findings at 7 T. *Radiology*, *261*(1), 199–209. https://doi.org/10.1148/radiol.11101651

Hogan, R. E., Mark, K. E., Wang, L., Joshi, S., Miller, M. I., & Bucholz, R. D. (2000). Mesial Temporal Sclerosis and Temporal Lobe Epilepsy: MR Imaging Deformation-based Segmentation of the Hippocampus in Five Patients. *Radiology*, *216*(1), 291–297. https://doi.org/10.1148/radiology.216.1.r00jl41291

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

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., & Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, *30*(9), 1617–1634. https://doi.org/10.1109/TMI.2011.2138152

Iglesias, J. E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., & Reuter, M. (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

Kennedy, D. N., Filipek, P. A., & Caviness, V. S. (1989). Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, *8*(1), 1–7. https://doi.org/10.1109/42.20356

Kerchner, G. A., Deutsch, G. K., Zeineh, M., Dougherty, R. F., Saranathan, M., & Rutt, B. K. (2012). Hippocampal CA1 apical neuropil atrophy and memory performance in Alzheimer's disease. *NeuroImage*, *63*(1), 194–202. https://doi.org/10.1016/j.neuroimage.2012.06.048

La Joie, R., Perrotin, A., de La Sayette, V., Egret, S., Doeuvre, L., Belliard, S., … Chételat, G. (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

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. https://doi.org/10.1017/CBO9781139087759

Machts, J., Vielhaber, S., Kollewe, K., Petri, S., Kaufmann, J., & Schoenfeld, M. A. (2018). Global Hippocampal Volume Reductions and Local CA1 Shape Deformations in Amyotrophic Lateral Sclerosis. *Frontiers in Neurology*, *9*. https://doi.org/10.3389/fneur.2018.00565

Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L., & Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging*, *31*(1), 192–203. https://doi.org/10.1002/jmri.22003

Marques, J. P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.-F., & Gruetter, R. (2010). MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *NeuroImage*, *49*(2), 1271–1281. https://doi.org/10.1016/j.neuroimage.2009.10.002

Marques, J. P., & Norris, D. G. (2018). How to choose the right MR sequence for your research question at 7T and above? *NeuroImage*, *168*, 119–140. https://doi.org/10.1016/j.neuroimage.2017.04.044

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. *Frontiers in Cellular Neuroscience*, *8*. https://doi.org/10.3389/fncel.2014.00095

Morey, R. D., & Rouder, J. N. (2019). BayesFactor: Computation of Bayes Factors for Common Designs. (Version 0.92) [R package]. Retrieved from https://cran.r-project.org/package=BayesFactor

Naidich, T. P., Delman, B. N., Fatterpekar, G. M., Gültekin, S. H., Aguinaldo, J. G. S., & Hof, P. R. (2003). Neuroanatomy at 9.4 Tesla: MR Microscopy of Formalin-Fixed Specimens of the Human Brain , Neuroa

natomy at 9.4 Tesla: MR Microscopy of Formalin-Fixed Specimens of the Human Brain. *Rivista Di Neuroradiologia*, *16*(2_suppl_part2), 164–166. https://doi.org/10.1177/1971400903016SP238

O'Brien, K. R., Kober, T., Hagmann, P., Maeder, P., Marques, J., Lazeyras, F., … Roche, A. (2014). Robust T1-Weighted Structural Brain Imaging and Morphometry at 7T Using MP2RAGE. *PLOS ONE*, *9*(6), e99676. https://doi.org/10.1371/journal.pone.0099676

Pipitone, J., Park, M. T. M., Winterburn, J., Lett, T. A., Lerch, J. P., Pruessner, J. C., … Chakravarty, M. M. (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

Pluta, J., Yushkevich, P., Das, S., & Wolk, D. (2012). In vivo analysis of hippocampal subfield atrophy in mild cognitive impairment via semi-automatic segmentation of T2-weighted MRI. *Journal of Alzheimer's Disease*, *31*(1), 85–99. https://doi.org/10.3233/JAD-2012-111931

R Core Team. (2019). R: A Language and environment for statistical computing. (Version 3.6.1). Retrieved from https://cran.r-project.org/.

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Sapolsky, R. M. (2001). Depression, antidepressants, and the shrinking hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(22), 12320–12322. https://doi.org/10.1073/pnas.231475998

Shaw, T. B., Bollmann, S., Atcheson, N. T., Guo, C., Fripp, J., Salvado, O., & Barth, M. (2019). Non-Linear Realignment Improves Hippocampus Subfield Segmentation. *BioRxiv*, 597856. https://doi.org/10.1101/597856

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175–200. https://doi.org/10.20982/tqmp.12.3.p175

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*. https://doi.org/10.1186/s12880-015-0068-x

The Jamovi Project. (2019). Jamovi (Version 0.9). Retrieved from https://www.jamovi.org

Tustison, N. J., Avants, B. B., Cook, P. A., Yuanjie Zheng, Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., … Avants, B. B. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage*, *99*, 166–179. https://doi.org/10.1016/j.neuroimage.2014.05.044

Tustison, N. J., Holbrook, A. J., Avants, B. B., Roberts, J. M., Cook, P. A., Reagh, Z. M., … Yassa, M. A. (2017). The ANTs Longitudinal Cortical Thickness Pipeline. *BioRxiv*, 170209. https://doi.org/10.1101/170209

Vincent, R. D., Neelin, P., Khalili-Mahani, N., Janke, A. L., Fonov, V. S., Robbins, S. M., … Evans, A. C. (2016). MINC 2.0: A Flexible Format for Multi-Modal Images. *Frontiers in Neuroinformatics*, *10*. https://doi.org/10.3389/fninf.2016.00035

Wang, D., & Doddrell, D. (2005). Geometric Distortion in Structural Magnetic Resonance Imaging. *Current Medical Imaging Reviews*, *1*(1), 49–60. https://doi.org/10.2174/1573405052953029

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., & Yushkevich, P. A. (2013). Multi-Atlas Segmentation with Joint Label Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 611–623. https://doi.org/10.1109/TPAMI.2012.143

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., Kuijf, H. J., Honingh, A. M., Wang, H., Pluta, J. B., Das, S. R., … Geerlings, M. I. (2016). Automated hippocampal subfield segmentation at 7 tesla MRI. *AJNR. American Journal of Neuroradiology*, *37*(6), 1050–1057. https://doi.org/10.3174/ajnr.A4659

Wu, G., Munsell, B. C., Zhan, Y., Bai, W., Sanroma, G., & Coupé, P. (2017). *Patch-Based Techniques in Medical Imaging: Third International Workshop, Patch-MI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Proceedings*. Springer.

Xie, L., Shinohara, R. T., Ittyerah, R., Kuijf, H. J., Pluta, J. B., Blom, K., … Wisse, L. E. M. (2018). Automated Multi-Atlas Segmentation of Hippocampal and Extrahippocampal Subregions in Alzheimer's Disease at 3T and 7T: What Atlas Composition Works Best? *Journal of Alzheimer's Disease: JAD*, *63*(1), 217–225. https://doi.org/10.3233/JAD-170932

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, *31*(3), 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje, E. C., … Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping*, *36*(1), 258–287. https://doi.org/10.1002/hbm.22627