Quantifying Numerical and Spatial Reliability of Amygdala and Hippocampal Subdivisions in FreeSurfer

Nicholas J. Buser[a], Christopher R. Madan[b], Jamie L. Hanson[a,*]

[a] *University of Pittsburgh, Pittsburgh, PA, USA*
[b] *University of Nottingham, Nottingham, UK*

Abstract

On-going, large-scale neuroimaging initiatives have produced many MRI datasets with hundreds, even thousands, of individual participants and scans. These databases can aid in uncovering neurobiological causes and correlates of poor mental health, disease pathology, and many other important factors. While volumetric quantification of brain structures can be completed by expert hand-tracing, automated segmentations are becoming the only truly tractable approach for particularly large datasets. Here, we assessed the spatial and numerical reliability for newly-deployed automated segmentation of hippocampal subfields and amygdala nuclei in FreeSurfer. In a sample of participants with repeated structural imaging scans (N=118), we found numerical reliability (as assessed by intraclass correlations) to be generally high, with 92% of the subregions having ICCs above 0.90 and the remainder still above 0.75. Spatial reliability was lower with only 11% of regions having Dice coefficients above 0.90, but 70% with Dice coefficients above 0.75. Of particular concern, three regions, the hippocampal fissure, the anterior amygdaloid area, and the paralaminar nucleus, had only moderate spatial reliability (0.50-0.75). We also examined correlations between spatial reliability and person-level factors (e.g., age, inter-scan interval, and difference in image quality). For these factors, inter-scan interval and image quality were related to variations in spatial reliability. Examined collectively, our work suggests strong numerical and spatial reliability for the majority of hippocampal and amygdala subdivisions; however, caution should be exercised for a few regions with more variable reliability.

1. Introduction

Continually clear from a large body of research is that the hippocampus and amygdala play key roles in emotion and stress-responding. Critically, both of these subcortical structures show volumetric alterations in different neurodegenerative diseases and various forms of psychopathologies, including Alzheimer's,

*Corresponding author
Email address: jamie.hanson@pitt.edu (Jamie L. Hanson)

Major Depression, Anxiety Disorders, and Autism. Continued study of these regions could be critical in understanding cognitive processes—such as memory, decision making, emotion—and may lead to novel intervention strategies for different disorders.

Early studies focused on the hippocampus and amygdala typically examined volumes of these regions using expert manual tracing [Gunten et al., 2000, Gunten and Ron, 2004, MacQueen et al., 2003, Yucel et al., 2007]. These approaches were often exceedingly time-intensive, but were necessary to obtain reliable and valid measures of the size of these key brain areas. As work in this space continues, large-scale structural MRI-datasets (Ns from 100s to 1000s) are now more commonly-available and work has shifted from manual tracing of regional volumes, and has instead leveraged ever-improving computational algorithms to automatically segment structural images into their component anatomical structures [**?**]. These approaches represent a scalable and easy method to potentially test relations between volumetric measures of these two structures and psychological variables of interest.

A commonly-used software suite, FreeSurfer [Fischl, 2012] provides a host of functions for structural MRI processing and analysis, including segmenting subcortical structures. Past work has examined both validity and reliability of hippocampus and amygdala segmentation in FreeSurfer [Jovicich et al., 2009, Madan and Kensinger, 2017, Morey et al., 2010]. One can think of validity as how well an output aligns with "ground-truth" (e.g., comparing FreeSurfer automated amygdala segments to expertly hand-traced volumes), while reliability is about consistency of outputs (e.g., comparing FreeSurfer automated amygdala from repeated scans of the same person, without consideration of any "ground-truth"). Previous works have found strong reliability for FreeSurfer, in terms of hippocampus and amygdala segmentations. Published reports examining test-retest reliability of subcortical volume measures have noted intraclass correlations from FreeSurfer ranging from 0.977-0.987 for the hippocampus and 0.806-0.889 for the amygdala [Jovicich et al., 2013, Liem et al., 2015, Wonderlick et al., 2009]. In regard to validity, results have been more mixed. Further work has investigated validity by comparing the spatial and numeric overlap between the volumes produced by FreeSurfer against those produced by expert hand tracing, finding relatively high correlations and Dice coefficients for the hippocampus but lower performance on the amygdala (hippocampus: r=0.82, Dice coefficient=0.82 ; amygdala: r=0.56, Dice coefficient=0.72) [Hanson et al., 2012, Morey et al., 2009].

In considering both the hippocampus and amygdala, each of these brain structures are often discussed as unitary structures; however, a large body of basic molecular and cognitive neuroscience research underscores that the hippocampus and amygdala each consist of multiple distinct subregions with different information-processing roles. For example, the hippocampus can be subdivided into: Dentate Gyrus, critical for pattern separation [Neunuebel and Knierim, 2014]; Cornu Ammonis (CA) 3, central to pattern completion [Guzman et al., 2016]; CA 1, important for input integration from CA3 and entorhinal cortex [Bittner et al., 2015]; and Subiculum being relevant for memory

retrieval [Roy et al., 2017]. The majority of past structural neuroimaging work has integrated over all of these regions and functions, using measures of whole hippocampal volume. This may mean a loss of specificity in regards to basic cognitive processes, as well as neurobiological alterations seen in different disorders. By examining subcortical structure at a more fine-grain scale, results can be more precisely fit to their root cause and better interpreted in light of their theoretical implications.

Responding to this issue, the developers of FreeSurfer have expanded their segmentation methods to include a more granular segmentation of hippocampal subregions [Iglesias et al., 2016]. To do this, they used ultra-high-resolution T1-weighted scans of post-mortem samples, segmented subfields of the hippocampus by hand, and then used these to develop an automated algorithm. Mirroring the whole-hippocampus reliability work, there appears to be good numeric reliability and slightly lower spatial reliability. Numeric reliability is focused on the consistent overall volume size (as indexed by the number of voxels in a region), whereas spatial reliability entails that the set of voxels classified are the same across both cases. These forms of reliability are typically correlated, but segments could have high numeric reliability but low spatial reliability. In such a case, the same number of voxels are being labelled as a brain region, but the voxels are actually spatially divergent (and may not be the same brain region). Past work has observed high numeric and moderately high spatial reliability for the hippocampal subfields, reporting ICCs ranging from 0.70 to 0.97 and Dice coefficients ranging from approximately 0.60-0.90 [Brown et al., 2020, Whelan et al., 2016].

The amygdala, likewise, has its own subdivisions and the reliability of these subdivisions are still unclear. The FreeSurfer team similarly expanded their segmentation pipeline to cover a set of subdivisions for the amygdala. The algorithm they employ is trained on manually segmented amygdala nuclei from high-definition 7 Tesla ex-vivo MR images and divides this structure into 9 labelled sub-regions. They applied this segmentation to datasets looking at populations with autism (ABIDE) [Martino et al., 2013] and those at risk for Alzheimer's disease (ADNI) [Jack et al., 2008] finding significant improvements in condition detection when this more fined grained view of the amygdala was used in the model [Saygin et al., 2017]. However, critically omitted from this work were direct assessment of numeric and spatial reliability for amygdala subdivisions. This is a significant omission, as it is currently unclear whether this fine-grained segmentation is consistent in the areas it is automatically dividing and outputting. Such gaps are important to fill-in, given that many groups are using these algorithms for applied purposes and reporting differences between clinical and non-clinical populations [Morey et al., 2020, Zheng et al., 2019].

Motivated by these facts, here, we seek to provide an in-depth examination of reliability, both numerically and spatially, for FreeSurfer derived hippocampal and amygdala subdivisions. We will leverage a public-access dataset of repeated structural scans with a reasonable sample size (n=118). In addition to this first-order goal, we will also consider whether person-level (e.g., age, sex) and MR-acquisition (e.g., interval between repeated scans; MRI quality) factors

influence the reliability of these subdivisions. Of note, recent work suggests that MR quality can significantly drive signal variations in structural MRI analyses [Gilmore et al., 2019, Madan and Kensinger, 2017]. Pursuing these aims can inform whether all subdivisions are truly "reliable" and should be explored in FreeSurfer-related analyses, or if caution should be taken in morphometric comparisons (especially for those working in applied areas, i.e., tests of amygdala subdivisions in depressed vs. non-depressed groups).

2. Methods

2.1. Participants

Data from an open-access neuroimaging initiative, the Consortium for Reliability and Reproducibility (CoRR) [Zuo et al., 2014], was used to investigate numeric and spatial reliability of FreeSurfer's amygdala and hippocampal subregion segmentation algorithms. CoRR includes structural and functional neuroimaging scans from participants, repeating scans across a short time-scale to see the stability of MRI-based metrics. For this work, T1-weighted MR images and demographic data were leveraged from two CoRR-related projects at Beijing Normal University (BNU).

The first study employed a test-retest interval design for 57 participants recruited from BNU's campus with an interval between scans of 41 +/- 5 days [Lin et al., 2015]. The second employed a similar design, but with 61 participants and a longer interval between scans of 161 +/- 15 days [Huang et al., 2016]. A total of 118 participants were studied across these two projects with 55 male participants (46.7%). The average age was 21.3 years old with a standard deviation of 1.9 years. The interval between scans ranged from 33 and 189 days, with an average of 103 and a standard deviation of 61 days. Of note, there was a significant difference in inter-scan interval across the two sites ($p<.005$) with one site (BNU-2) having longer intervals (>3 months). This allowed us to investigate a wider-range of inter-scan intervals as an important variable of interest.

2.2. MRI Scan Parameters

MR images were acquired with a Siemens TrioTrim 3T scanner at Beijing Normal University. T1-weighted MR images were acquired using a 3D-MPRAGE sequence (TR: 2.53 s; TE: 3.39 ms; flip angle: 7.0 degrees; slice thickness: 1.3 mm). Of note, there were slight variations in acquisition parameters; these are noted and discussed in the supplemental material.

2.3. Structural Neuroimaging Processing (Freesurfer)

Standard-processing approaches from FreeSurfer (e.g., cortical reconstruction; volumetric segmentation) were performed in version 6.0 (with in-development routines for the hippocampal and amygdala subdivisions) on a Linux RedHat

computing cluster. FreeSurfer is a widely-documented and freely available morphometric processing tool suite (http://surfer.nmr.mgh.harvard.edu). The technical details of these procedures are described in prior publications [Fischl, 2012, Dale et al., 1999, Fischl et al., 2004, 1999a,b]. Briefly, this processing includes motion correction and intensity normalization of T1-weighted images, removal of non-brain tissue using a hybrid watershed/surface deformation procedure [Ségonne et al., 2004], automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles), tessellation of the gray matter white matter boundary, and derivation of cortical thickness. Scans from two subjects failed to run to completion in this pipeline and both subjects were removed from further analysis.

Given our scans were repeated acquisitions from the same participants, we processed our T1-weighted scans in FreeSurfer's longitudinal stream [Reuter et al., 2012]. For this approach, an unbiased within-subject template image was created using robust, inverse consistent registration [Reuter et al., 2010]. Several processing steps, such as skull stripping, Talairach transforms, atlas registration, as well as spherical surface maps and parcellations, were then initialized with common information from the within-subject template, significantly increasing reliability and statistical power [Reuter et al., 2012]. Following this longitudinal pipeline, a third script of operations was completed to further segment the hippocampus and amygdala into their subnuclei. The longitudinal pipeline described above augments this additional script, allowing the algorithm to make use of the scans at both timepoints to jointly compute segmentation through Bayesian inference [Iglesias et al., 2016].

The hippocampal segmentation method [Iglesias et al., 2015] is based on a hippocampal atlas initially produced from a dataset of 15 hand-traced high definition ex-vivo T1-weighted 7 T scans then applied to a set of 39 standard resolution in-vivo MPRAGE scans using parameterized mesh deformations and a probabilistic atlas classification approach. This atlas is used for algorithmic segmentation of MR images pre-processed through the FreeSurfer recon-all pipeline. These images were classified using a parameterized generative model and optimizing the likelihood that any given voxel belongs to the label of a particular hippocampal region in a Bayesian inference framework. For additional information, see [Iglesias et al., 2015]. The atlas for this method partitions the hippocampus into the following 13 subfields: (1) Parasubiculum, (2) Presubiculum, (3) Subiculum, (4) CA1, (5) CA2/3,(6) CA4, (7) Granule Cell and Molecular Layer of Dentate Gyrus (GC-ML-DG), (8) Molecular layer, (9) Fimbria, (10) Hippocampal Fissure, (11) Hippocampus-Amygdala-Transition-Area (HATA), (12) Hippocampal Tail, and (13) Alveus.

For the amygdala, the automated segmentation method [Saygin et al., 2017] is based on an atlas produced from 10 hand-traced high definition ex-vivo T1w 7 T scans (5 participants traced bilaterally). As in the hippocampal atlas, this manually segmented ex-vivo data was then applied to the probabilistic classification of the nodes on a parameterized deformation mesh of the amygdala. Similar to the hippocampus, the segmentation of later input data is performed

in the framework of Bayesian inference. The amygdala atlas partitions the structure into the following 8 subnuclei: (1) Anterior, (2) Corticoamygdaloid Transition Area, (3) Basal, (4) Lateral, (5) Accessory Basal, (6) Central, (7) Cortical Medial, and (8) Paralaminar. These methods (for both the hippocampal and amygdala subdivisions) were applied to both scans (test and retest) for each participant. For both the hippocampal subfields and amygdala nuclei, volume (number of voxels) for each subdivision was extracted and used in numeric reliability analysis. Spatial information (labelled voxels in axial, coronal, and spatial orientations) was also output for each subdivision, for use in spatial reliability analysis.

2.3.1. MRI (Automated) Quality Assessment

The Computational Anatomy Toolbox 12 (CAT12) toolbox from the Structural Brain Mapping group, implemented in SPM12, was used to generate a quantitative metric indicating the quality of each collected MR image [Gaser and Kurth, 2017]. The method employed considers four summary measures of image quality: (1) noise to contrast ratio, (2) coefficient of joint variation, (3) inhomogeneity to contrast ratio, and (4) root mean squared voxel resolution. To produce a single aggregate metric that serves as an indicator of overall quality, this toolbox normalizes each measure and combines them using a kappa statistic-based framework, for optimizing a generalized linear model through solving least squares [Dahnke et al., 2015]. After extracting a single quality metric for each scan we subtracted the score for each participant's first scan from their second, using that as a measure of interscan quality difference.

2.4. Derivation of Reliability Measures

To assess the reliability of the numeric volume output for hippocampus and amygdala subdivisions, we computed intraclass correlation coefficients (ICC) between each labelled sub-region for the test and the retest MRI scans. Of note, an ICC is a descriptive statistic indicating the degree of agreement between two (or more) sets of measurements. The statistic is similar to a bivariate correlation coefficient insofar as it has a range from 0-1 and higher values represent a stronger relationship. An ICC, however, differs from the bivariate correlation in that it works on groups of measurements and gives an indication of the numeric cohesion across the given groups [McGraw and Wong, 1996]. The ICC was calculated separately for each sub-region using the statistical programming language R, with the icc function from the package irr [Gamer and Lemon, 2012]. A two-way model with absolute agreement was used in order to investigate the consistency of the classification on the randomly sampled subject pool. Although there are no definitive guidelines for precise interpretation of ICCs, results have frequently been binned into three (or four) quality groups where 0.0-0.5 is "poor", 0.50-0.75 is "moderate", 0.75-0.9 is "good" and 0.9-1.0 is "excellent" [Cicchetti, 1994, Koo and Li, 2016].

Although ICCs serve as an indication of numeric reliability, it is only one metric and may be incomplete, particularly when we think about the spatial

information present in MRI volumes. Indeed, even with numeric similarity, there may be discrepancies in the specific spatial voxels labeled for a given subdivision. To assess whether the voxels assigned to each region were the same between the two timepoints, we calculated the Sørensen-Dice Coefficient using the @DiceMetric program in the AFNI fMRI software package [Cox, 1996]. The Dice coefficient is calculated as $\frac{(2TP)}{(2TP+FP+FN)}$ [TP = True Positive; FP = False Positive; FN = False Negative] and gives an equal weight to criteria of positive predictive value and sensitivity in assessing spatial reliability of subdivisions.

2.4.1. Statistical Analysis

Once reliability was calculated, we examined person-level (e.g., age, sex) and MR-acquisition (e.g., interval between repeated scans; MRI quality) factors in relation to these measures. Many different factors may impact amygdala and hippocampal segmentation. For example, past work suggests volumes (of the hippocampus and amygdala) vary with participant age and sex; this association is particularly strong for the hippocampus [Daugherty et al., 2016, Nobis et al., 2019] and suggestive data similarly for the amygdala [Malykhin et al., 2008, Marwha et al., 2017, Perlaki et al., 2014, Pressman et al., 2016]. Although the interscan intervals here are short, the longer one waits between scans the more likely it is that actual brain volumes could change, so we considered the length of times between scans as a possible predictor. Finally, image quality has been shown to have a significant effect on brain volume measurements [Gilmore et al., 2019]. Noisier image may lead to gray/white matter misclassification, and impact reliability between different scans. To consider these potential effects, we examined each region's reliability in relation to age, sex, interscan interval, and difference in the CAT12 quality metric. We constructed regression models to examine connections between reliability metrics, and person-level and MR-acquisition factors. Of note, the difference in quality between the two scans (scan-2 minus scan-1) was in these analyses.

3. Results

3.1. Hippocampus Reliability

Using ICC analysis, we found consistently high levels of numeric reliability within the hippocampus. Of note here is that while the region remained in a highly reliable range, the hippocampal fissure performed substantially worse than the remaining areas, with a more pronounced effect in the left hemisphere but with scores from both hemispheres being the only ones with values under 0.90 (see Table 1). Using Dice coefficients as metrics of spatial reliability, we found largely reassuring results across all hippocampal regions, with again, a substantial drop in reliability for the hippocampal fissure, with scores at 0.53 and 0.58 for the left and right hemisphere respectively. These are reported with numeric reliability (see Table 1). Of note, this data aggregates across the two scanning sites; however, these reliability metrics were highly correlated when calculated independently for each site (r=.99).

region	icc lh	icc rh	dice lh	dice rh
Hippocampal tail	0.96	0.94	0.91	0.91
subiculum.body	0.96	0.97	0.88	0.89
CA1.body	0.95	0.96	0.77	0.82
subiculum.head	0.97	0.98	0.84	0.85
hippocampal.fissure	0.78	0.87	0.53	0.58
presubiculum.head	0.94	0.97	0.86	0.86
CA1.head	0.97	0.98	0.84	0.86
presubiculum.body	0.90	0.94	0.88	0.86
parasubiculum	0.93	0.97	0.79	0.80
molecular layer HP.head	0.97	0.96	0.74	0.76
molecular layer HP.body	0.94	0.96	0.72	0.75
GC.ML.DG.head	0.95	0.96	0.70	0.71
CA3.body	0.97	0.96	0.72	0.75
GC.ML.DG.body	0.90	0.94	0.70	0.72
CA4.head	0.94	0.95	0.85	0.85
CA4.body	0.91	0.95	0.86	0.87
fimbria	0.91	0.92	0.75	0.76
CA3.head	0.93	0.95	0.72	0.71
HATA	0.95	0.95	0.78	0.79

Table 1: ICCs for Hippocampal Subfields. Color coding is in accordance with excellent-good-poor scores with 0.90-1.00 constituting excellent scores in green, 0.75-0.90 constituting good scores in yellow, and 0.00-0.75-constituting poor scores in red.

region	icc lh	icc rh	dice lh	dice rh
Lateral.nucleus	0.97	0.98	0.93	0.93
Basal.nucleus	0.95	0.97	0.92	0.92
Accessory.Basal.nucleus	0.96	0.97	0.87	0.88
Anterior.amygdaloid.area.AAA	0.92	0.92	0.61	0.62
Central.nucleus	0.92	0.92	0.86	0.87
Medial.nucleus	0.87	0.89	0.87	0.87
Cortical.nucleus	0.95	0.94	0.87	0.88
Corticoamygdaloid.transitio	0.97	0.98	0.75	0.76
Paralaminar.nucleus	0.96	0.96	0.57	0.57

Table 2: Intraclass Correlation Coefficients (ICCs) for Amygdala Subnuclei. Color coding is in accordance with excellent-good-poor scores with 0.90-1.00 constituting excellent scores in green, 0.75-0.90 constituting good scores in yellow, and 0.00-0.75-constituting poor scores in red.

3.2. Amygdala Reliability

Within the amygdala the numeric reliability was all around very high (<.87). Values are reported (see Table 2). While most of the subnuclei in the amygdala were high in spatial reliability as well, two regions stood out as potentially problematic: the Anterior amygdaloid area and the Paralaminar nucleus, both with average dice coefficients in the 0.50-0.70 range. This poor spatial reliability is graphically depicted in Figure 1, with the Paralaminar nucleus segments noted for the first and second scan from an exemplar subject. For contrast, a region, the Lateral amygdala nucleus, with good spatial reliability is shown in Figure 2. Spatial reliability values for all subdivisions are reported in Table 2. Again, of note, this data aggregates across the two scanning sites; however, these reliability metrics were highly correlated when calculated independently for each site (r=.99).

3.3. Reliability Differences in Relation to Person-level and MR-acquisition Factors

Given the generally high numeric reliability for all subdivisions, we only examined associations between spatial reliability and subject-level variables. Correlations between the Hippocampal-subfield dice coefficients and our subject-level variables are reported in Table 3. Length between neuroimaging sessions (test and retest) had a significant negative correlation with dice-coefficients for all regions. Differences in image quality had negative significant relationships for many, but not all, subfields. Finally, for a small number of regions, sex and age were related to spatial reliability; however, these variables were not significantly related to spatial reliability for most regions.

The correlations for the amygdala nuclei are reported in Table 4. For these regions as well, the relationship with length between test and retest was negative and significant. The effect of difference in image quality was also negative and significant for most but not all regions. Finally, several regions reported a

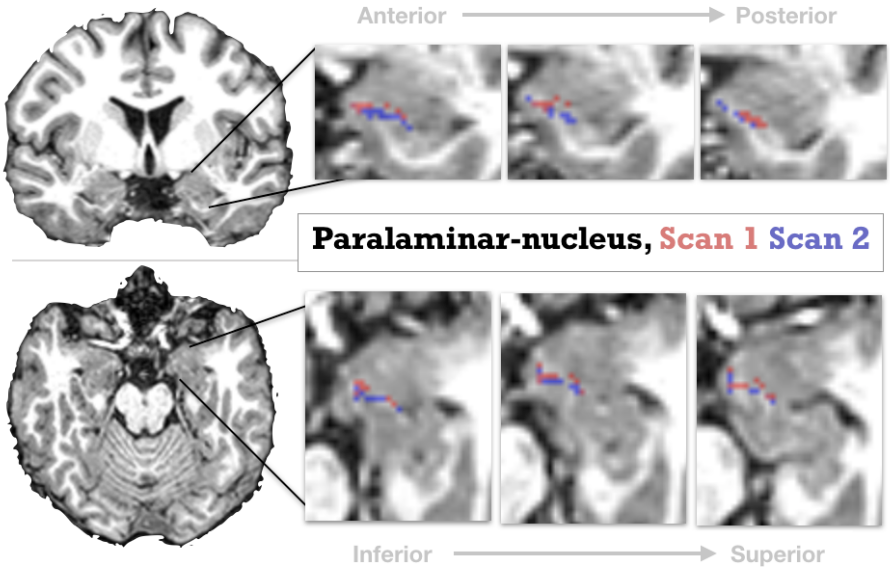

Figure 1: Graphic depictions showing coronal (top) and axial (bottom) views with magnified depictions of the segments of the Paralaminer nucleus from the repeated scans (Scan 1 shown in Red; Scan 2 shown in Purple). The anatomical (T1w) image underlaid is the unbiased subject template from an example participant. In the top panel, multiple slices are depicted left to right, moving anterior to posterior; in the bottom panel, slices move left to right from inferior to superior.

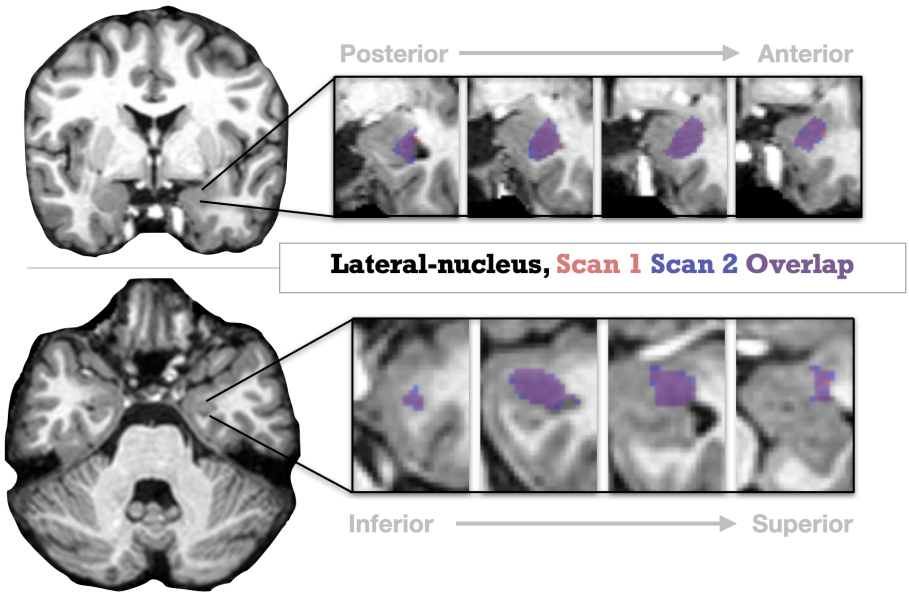

Figure 2: Graphic depictions showing coronal (top) and axial (bottom) views with magnified depictions of the segments of the Lateral nucleus from the repeated scans (Scan 1 shown in Red; Scan 2 shown in Purple). The anatomical (T1w) image underlaid is the unbiased subject template from an example participant. In the top panel, multiple slices are depicted left to right, moving anterior to posterior; in the bottom panel, slices move left to right from inferior to superior.

regions	sex lh	sex rh	ISI lh	ISI rh	age lh	age rh	qc lh	qc rh
parasubiculum	0.22	0.01	-0.28	-0.36	0.09	0.04	-0.21	-0.20
HATA	0.09	0.21	-0.39	-0.43	0.14	0.25	-0.25	-0.31
fimbria	0.11	0.06	-0.31	-0.41	0.13	0.11	-0.07	-0.23
hippocampal.fissure	0.04	0.01	-0.49	-0.54	0.14	0.16	-0.27	-0.32
Hippocampal tail	0.12	0.05	-0.29	-0.32	0.15	0.19	-0.29	-0.30
presubiculum.head	0.13	-0.03	-0.27	-0.43	0.16	0.06	-0.19	-0.16
presubiculum.body	0.11	0.13	-0.26	-0.34	0.12	0.10	-0.15	-0.21
subiculum.head	0.09	0.09	-0.35	-0.31	0.17	0.07	-0.21	-0.19
subiculum.body	0.12	0.02	-0.29	-0.33	0.10	0.09	-0.15	-0.25
CA1.head	0.15	0.11	-0.44	-0.48	0.21	0.20	-0.21	-0.29
CA1.body	0.05	0.03	-0.30	-0.34	0.07	0.09	-0.23	-0.31
CA3.head	0.13	-0.04	-0.39	-0.52	0.17	0.20	-0.11	-0.23
CA3.body	0.08	0.04	-0.32	-0.40	0.07	0.07	-0.11	-0.08
CA4.head	0.12	0.02	-0.26	-0.44	0.13	0.15	0.01	-0.22
CA4.body	0.09	0.05	-0.22	-0.31	0.08	0.07	-0.06	-0.16
GC.ML.DG.head	0.16	0.02	-0.34	-0.50	0.17	0.15	-0.13	-0.28
GC.ML.DG.body	0.09	0.03	-0.30	-0.42	0.10	0.05	-0.12	-0.22
molecular layer HP.head	0.17	0.10	-0.38	-0.44	0.20	0.11	-0.15	-0.21
molecular layer HP.body	0.09	0.01	-0.31	-0.41	0.09	0.02	-0.12	-0.21

Table 3: Correlation coefficient for the bivariate regressions between Hippocampal Subfield Dice Coefficients and subject-level covariates: sex, interscan interval (ISI), age, and difference in qc score (qc).

regions	sex lh	sex rh	ISI lh	ISI rh	age lh	age rh	qc lh	qc rh
Lateral.nucleus	0.08	0.07	-0.30	-0.41	0.11	0.14	-0.16	-0.23
Basal.nucleus	0.09	0.12	-0.28	-0.35	0.09	0.13	-0.17	-0.21
Central.nucleus	0.13	0.17	-0.31	-0.35	0.09	0.12	-0.22	-0.24
Medial.nucleus	0.13	0.18	-0.30	-0.34	0.08	0.13	-0.22	-0.23
Cortical.nucleus	0.14	0.21	-0.29	-0.33	0.08	0.12	-0.21	-0.22
Accessory.Basal.nucleus	0.14	0.21	-0.28	-0.32	0.07	0.12	-0.20	-0.22
Corticoamygdaloid.transitio	0.13	0.21	-0.32	-0.36	0.08	0.17	-0.20	-0.23
Anterior.amygdaloid.area.AAA	0.07	0.09	-0.26	-0.23	0.03	0.01	-0.16	-0.10
Paralaminar.nucleus	0.01	0.01	-0.35	-0.30	0.04	0.06	-0.18	-0.07

Table 4: Correlation coefficient for the bivariate regressions between Amygdala Subnuclei Dice Coefficients and subject-level covariates: sex, interscan interval (ISI), age, and difference in qc score (qc).

positive relationship with sex, yet none displayed a relationship on the basis of age.

4. Discussion

In this paper, we assessed the numeric and spatial reliability of FreeSurfer's hippocampus and amygdala sub-segmentation algorithms. The ICC's, serving as our indicator of numeric reliability, were overall very high (hippocampal subfields: 0.78-0.98; amygdala nuclei: 0.87-0.98), indicating that FreeSurfer is generally numerically reliable in providing overall volume for each subregion. The dice coefficients, serving as our indicator of spatial reliability, were also generally high, though slightly lower than the ICCs. Of note and potential concern, a few subdivisions had fairly low spatial reliability, suggesting unreliable segmentation.

While these results suggest that, for the most part, volumetric outputs of amygdala and hippocampal subdivisions are reliable, the drop in spatial reliability may mean researchers should exercise caution in the analysis and interpretation of these specific areas. For example, values from the hippocampal fissure, the anterior amygdaloid area, and the paralaminar nucleus showed only moderate spatial reliability through their dice coefficients (hippocampal fissure: 0.53/0.58, anterior amygdaloid l/h: 0.61/0.62, paralaminar l/h: 0.57/0.57). Because the spatial reliability of these three areas is relatively poor, studies which interpret changes in volume within or across subjects might be using segmentations which contain improperly (or inconsistently) classified voxels within those regions. Several studies have already reported significant findings from the paralaminar region, given the questionable reproducibility of its anatomical bounding these findings may require further verification [Morey et al., 2020, Zheng et al., 2019]. The exact causes of this poor spatial reliability are currently unclear. Many of these areas are particularly small in size, meaning modest amounts of noise and error (e.g., head motion and field inhomogeneities) can

greatly influence spatial overlap. Therefore, and related, registration may be particularly challenging, leading to these decrements in spatial reliability.

We have extended the work of previous test-retest publications, applying similar approaches to the problem of testing the reliability of FreeSurfer's amygdala segmentation algorithm and replicating the findings which had already considered such reliability measures on the existing FreeSurfer methods for hippocampal subfield segmentation. In line with these previous publications we have found generally very high numeric reliability and overall moderate-to good indicators of spatial reliability [Schoemaker et al., 2016]. Of note, there is no previous work, to our knowledge, which characterizes this performance for the amygdala nuclei.

Correlation analyses indicated consistently negative relationships between Dice coefficients and the interval between scans. While a short time frame makes it unlikely, true changes in subcortical volume do occur over time and the longer the interval between scans, the more likely it is that detected changes are veridical. A number of regions showed significant drops in reliability associated with higher differences in image quality between time points. This further emphasizes the importance of considering the effect of image quality on statistical results. Only a small number of regions displayed sensitivity to changes in sex or age, suggesting that these demographic variables may not be driving major differences in the algorithm's reliability.

Of importance and as noted in our introduction, our findings only speak to the reliability of these measures, and not the validity of these segments. Future work for rigorous establishment of such algorithmic validity should be pursued to increase confidence in the overall performance of the FreeSurfer segmentations. Previous work has compared FreeSurfer's hippocampal subfields to hand drawn volumes [Herten et al., 2019, Worker et al., 2018], however, to our knowledge there is yet to be any comparison of the amygdala nuclei segmentation to such a gold standard. Reliable methods exist for expert manual segmentation of the amygdala [Entis et al., 2012]. A study which looks at the degree of overlap between such methods and the FreeSurfer algorithm for amygdala segmentation would be helpful for effective evaluation of validity.

In considering our results, it is important to note a few potential limitations of our work. First, the sample is a rather homogenous group of individuals and may not represent the greater population. All participants were recruited from Beijing Normal University and were neurotypical young adults. Due to this demographic, it is uncertain to what degree these findings will generalize outside of that domain and further research should be done to ascertain this. Additional work considering reliability of the method in a diverse set of populations (e.g., pediatric, elderly, mild cognitive impairment) would be helpful in ascertaining how well these findings generalize outside of our sample population. Given the variable levels of developmental stages in pediatric patients and the variability introduced through cognitive decline it is important to directly study the performance of these techniques developed directly on a normative population for these other contexts. Second, we do not have detailed information regarding more fine-grain drivers of differences in image quality. Particular types of mo-

tion, scanner inhomogeneities, or other such artifact-inducing conditions were all grouped into a single report of image quality. Given that subtle differences in registration and participant differences in movement, signal susceptibility, and so forth all have the ability to affect the performance of these algorithms it is possible that with a more detailed dataset regarding precise movement records or physiological information we might be able to better characterize the scan-time conditions which drive algorithmic performance differences related to overall image quality.

Those limitations notwithstanding, our work extends the information provided by previous publications regarding the reliability of FreeSurfer's subcortical segmentation for the hippocampus, amygdala, and their respective subregions. To our knowledge, this is the first work to directly investigate the test-retest reliability of the amygdala nuclei algorithm. The strengths of our work include a reasonably large sample size, the use of FreeSurfer's more robust longitudinal pipeline, and the report of mathematically rigorous measures of reliability. Our work provides additional confidence in interpreting those regions with high reliability and a necessary caution in interpretation of those with poorer results.

References

Katie C. Bittner, Christine Grienberger, Sachin P. Vaidya, Aaron D. Milstein, John J. Macklin, Junghyup Suh, Susumu Tonegawa, and Jeffrey C. Magee. 2015. Conjunctive input processing drives feature selectivity in hippocampal CA1 neurons. *Nature Neuroscience* 18, 8 (Aug. 2015), 1133–1142. https://doi.org/10.1038/nn.4062 Publisher: Nature Publishing Group.

Emma M. Brown, Meghan E. Pierce, Dustin C. Clark, Bruce R. Fischl, Juan E. Iglesias, William P. Milberg, Regina E. McGlinchey, and David H. Salat. 2020. Test-retest reliability of FreeSurfer automated hippocampal subfield segmentation within and across scanners. *NeuroImage* 210 (April 2020). https://doi.org/10.1016/j.neuroimage.2020.116563 Publisher: Academic Press Inc.

Domenic V. Cicchetti. 1994. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment* 6, 4 (1994), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Robert W. Cox. 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research* 29, 3 (1996), 162–173. https://doi.org/10.1006/cbmr.1996.0014 Publisher: Academic Press Inc.

Robert Dahnke, Gabriel Ziegler, Julian Grosskreutz, and Christian Gaser. 2015. Quality Assurance in Structural MRI. *21st Annual Meeting of the Organization For Human Brain Mapping* (2015), 1556.

Anders M. Dale, Bruce Fischl, and Martin I. Sereno. 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage* 9, 2 (1999), 179–194. https://doi.org/10.1006/nimg.1998.0395 Publisher: Academic Press Inc.

Ana M. Daugherty, Andrew R. Bender, Naftali Raz, and Noa Ofen. 2016. Age differences in hippocampal subfield volumes from childhood to late adulthood. *Hippocampus* 26, 2 (Feb. 2016), 220–228. https://doi.org/10.1002/hipo.22517 Publisher: John Wiley and Sons Inc.

Jonathan J. Entis, Priya Doerga, Lisa Feldman Barrett, and Bradford C. Dickerson. 2012. A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution MRI. *NeuroImage* 60, 2 (April 2012), 1226–1235. https://doi.org/10.1016/j.neuroimage.2011.12.073 Publisher: NIH Public Access.

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

Bruce Fischl, David H. Salat, André J.W. Van Der Kouwe, Nikos Makris, Florent Ségonne, Brian T. Quinn, and Anders M. Dale. 2004. Sequence-independent segmentation of magnetic resonance images. *NeuroImage* 23, SUPPL. 1 (2004). https://doi.org/10.1016/j.neuroimage.2004.07.016

Bruce Fischl, Martin I. Sereno, Roger B.H. Tootell, and Anders M. Dale. 1999a. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping* 8, 4 (1999), 272–284. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<272::AID-HBM10>3.0.CO;2-4

Bruce Fischl, Martin I. Sereno, Roger B.H. Tootell, and Anders M. Dale. 1999b. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping* 8, 4 (1999), 272–284. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<272::AID-HBM10>3.0.CO;2-4

Matthias Gamer and Jim Lemon. 2012. Various Coefficients of Interrater Reliability and Agreement. (2012).

Christian Gaser and Florian Kurth. 2017. Manual computational anatomy toolbox-cat12. *Structural Brain Mapping Group at the Departments of Psychiatry and Neurology, University of Jena* (2017).

Alysha Gilmore, Nicholas Buser, and Jamie L. Hanson. 2019. Variations in Structural MRI Quality Impact Measures of Brain Anatomy: Relations with Age and Other Sociodemographic Variables. *bioRxiv* (March 2019), 581876. https://doi.org/10.1101/581876 Publisher: Cold Spring Harbor Laboratory.

Armin von Gunten, Nicholas C. Fox, Lisa Cipolotti, and Maria A. Ron. 2000. A Volumetric Study of Hippocampus and Amygdala in Depressed Patients With Subjective Memory Problems. *The Journal of Neuropsychiatry and Clinical Neurosciences* 12, 4 (2000), 493–498. https://doi.org/10.1176/jnp.12.4.493

Armin von Gunten and Maria A. Ron. 2004. Hippocampal Volume and Subjective Memory Impairment in Depressed Patients. *European Psychiatry* 19, 7 (2004), 438–440. https://doi.org/10.1016/j.eurpsy.2004.05.003

Segundo Jose Guzman, Alois Schlögl, Michael Frotscher, and Peter Jonas. 2016. Synaptic mechanisms of pattern completion in the hippocampal CA3 network. *Science* 353, 6304 (Sept. 2016), 1117–1123. https://doi.org/10.1126/science.aaf1836 Publisher: American Association for the Advancement of Science.

Jamie L. Hanson, Jung W. Suh, Brendon M. Nacewicz, Matthew J. Sutterer, Amelia A. Cayo, Diane E. Stodola, Cory A. Burghy, Hongzhi Wang, Brian B.

Avants, Paul A. Yushkevich, Marilyn J. Essex, Seth D. Pollak, and Richard J. Davidson. 2012. Robust Automated Amygdala Segmentation via Multi-Atlas Diffeomorphic Registration. *Frontiers in Neuroscience* 6 (2012). https://doi.org/10.3389/fnins.2012.00166

Annika Herten, Kerstin Konrad, Helga Krinzinger, Jochen Seitz, and Georg G. von Polier. 2019. Accuracy and bias of automatic hippocampal segmentation in children and adolescents. *Brain Structure and Function* 224, 2 (March 2019), 795–810. https://doi.org/10.1007/s00429-018-1802-2 Publisher: Springer Verlag.

Lijie Huang, Taicheng Huang, Zonglei Zhen, and Jia Liu. 2016. A test-retest dataset for assessing long-Term reliability of brain morphology and resting-state brain activity. *Scientific Data* 3 (March 2016). https://doi.org/10.1038/sdata.2016.16 Publisher: Nature Publishing Groups.

Juan Eugenio Iglesias, Jean C. Augustinack, Khoa Nguyen, Christopher M. Player, Allison Player, Michelle Wright, Nicole Roy, Matthew P. Frosch, Ann C. McKee, Lawrence L. Wald, Bruce Fischl, and Koen Van Leemput. 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 (2015), 117–137. https://doi.org/10.1016/j.neuroimage.2015.04.042 Publisher: Elsevier B.V.

Juan Eugenio Iglesias, Koen Van Leemput, Jean Augustinack, Ricardo Insausti, Bruce Fischl, and Martin Reuter. 2016. Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. (2016). https://doi.org/10.1016/j.neuroimage.2016.07.020

Clifford R. Jack, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L. G. Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S. DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P. Debbins, Adam S. Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W. Weiner and. 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): Mri Methods. *Journal of Magnetic Resonance Imaging* 27, 4 (2008), 685–691. https://doi.org/10.1002/jmri.21049

Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, and Deborah Blacker. 2009. Mri-Derived Measurements of Human Subcortical, Ventricular and Intracranial Brain Volumes: Reliability Effects of Scan Sessions, Acquisition Sequences, Data Analyses, Scanner Upgrade, Scanner Vendors and Field Strengths. *NeuroImage* 46, 1 (2009), 177–192. https://doi.org/10.1016/j.neuroimage.2009.02.010

Jorge Jovicich, Moira Marizzoni, Roser Sala-Llonch, Beatriz Bosch, David Bartrés-Faz, Jennifer Arnold, Jens Benninghoff, Jens Wiltfang, Luca Roccatagliata, Flavio Nobili, Tilman Hensch, Anja Tränkner, Peter Schönknecht, Melanie Leroy, Renaud Lopes, Régis Bordet, Valérie Chanoine, Jean-Philippe Ranjeva, Mira Didic, Hélène Gros-Dagnac, Pierre Payoux, Giada Zoccatelli, Franco Alessandrini, Alberto Beltramello, Núria Bargalló, Olivier Blin, and Giovanni B. Frisoni. 2013. Brain Morphometry Reproducibility in Multi-Center 3t Mri Studies: a Comparison of Cross-Sectional and Longitudinal Segmentations. *NeuroImage* 83, nil (2013), 472–484. https://doi.org/10.1016/j.neuroimage.2013.05.007

Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (June 2016), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012 Publisher: Elsevier USA.

Franziskus Liem, Susan Mérillat, Ladina Bezzola, Sarah Hirsiger, Michel Philipp, Tara Madhyastha, and Lutz Jäncke. 2015. Reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly. *NeuroImage* 108 (March 2015), 95–109. https://doi.org/10.1016/j.neuroimage.2014.12.035 Publisher: Academic Press Inc.

Qixiang Lin, Zhengjia Dai, Mingrui Xia, Zaizhu Han, Ruiwang Huang, Gaolang Gong, Chao Liu, Yanchao Bi, and Yong He. 2015. A connectivity-based test-retest dataset of multi-modal magnetic resonance imaging in young healthy adults. *Scientific Data* 2 (Oct. 2015). https://doi.org/10.1038/sdata.2015.56 Publisher: Nature Publishing Groups.

G. M. MacQueen, S. Campbell, B. S. McEwen, K. Macdonald, S. Amano, R. T. Joffe, C. Nahmias, and L. T. Young. 2003. Course of Illness, Hippocampal Function, and Hippocampal Volume in Major Depression. *Proceedings of the National Academy of Sciences* 100, 3 (2003), 1387–1392. https://doi.org/10.1073/pnas.0337481100

Christopher R. Madan and Elizabeth A. Kensinger. 2017. Test–retest reliability of brain morphology estimates. *Brain Informatics* 4, 2 (June 2017), 107–121. https://doi.org/10.1007/s40708-016-0060-4 Publisher: Springer Berlin Heidelberg.

Nikolai V. Malykhin, Thomas P. Bouchard, Richard Camicioli, and Nicholas J. Coupland. 2008. Aging hippocampus and amygdala. *NeuroReport* 19, 5 (March 2008), 543–547. https://doi.org/10.1097/WNR.0b013e3282f8b18c

A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy,

C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham. 2013. The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism. *Molecular Psychiatry* 19, 6 (2013), 659–667. https://doi.org/10.1038/mp.2013.78

Dhruv Marwha, Meha Halari, and Lise Eliot. 2017. Meta-analysis reveals a lack of sexual dimorphism in human amygdala volume. *NeuroImage* 147 (Feb. 2017), 282–294. https://doi.org/10.1016/j.neuroimage.2016.12.021 Publisher: Academic Press Inc.

Kenneth O. McGraw and S. P. Wong. 1996. Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods* 1, 1 (1996), 30–46. https://doi.org/10.1037/1082-989X.1.1.30 Publisher: American Psychological Association Inc.

Rajendra A. Morey, Emily K. Clarke, Courtney C. Haswell, Rachel D. Phillips, Ashley N. Clausen, Mary S. Mufford, Zeynep Saygin, Mira Brancu, Jean C. Beckham, Patrick S. Calhoun, Eric Dedert, Eric B. Elbogen, John A. Fairbank, Robin A. Hurley, Jason D. Kilts, Nathan A. Kimbrel, Angela Kirby, Christine E. Marx, Scott D. McDonald, Scott D. Moore, Jennifer C. Naylor, Jared Rowland, Cindy Swinkels, Steven T. Szabo, Katherine H. Taber, Larry A. Tupler, Elizabeth E. van Voorhees, Ruth E. Yoash-Gantz, H. Ryan Wagner, and Kevin S. LaBar. 2020. Amygdala Nuclei Volume and Shape in Military Veterans With Posttraumatic Stress Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 5, 3 (March 2020), 281–290. https://doi.org/10.1016/j.bpsc.2019.11.016 Publisher: Elsevier Inc.

Rajendra A. Morey, Christopher M. Petty, Yuan Xu, Jasmeet Pannu Hayes, H. Ryan Wagner, Darrell V. Lewis, Kevin S. LaBar, Martin Styner, and Gregory McCarthy. 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *NeuroImage* 45, 3 (April 2009), 855–866. https://doi.org/10.1016/j.neuroimage.2008.12.033 Publisher: Academic Press.

Rajendra A. Morey, Elizabeth S. Selgrade, Henry Ryan Wagner, Scott A. Huettel, Lihong Wang, and Gregory McCarthy. 2010. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Human Brain Mapping* 31, 11 (Nov. 2010), 1751–1762. https://doi.org/10.1002/hbm.20973

Joshua P. Neunuebel and James J. Knierim. 2014. CA3 retrieves coherent representations from degraded input: Direct evidence for CA3 pattern completion

and dentate gyrus pattern separation. *Neuron* 81, 2 (Jan. 2014), 416–427. https://doi.org/10.1016/j.neuron.2013.11.017

Lisa Nobis, Sanjay G. Manohar, Stephen M. Smith, Fidel Alfaro-Almagro, Mark Jenkinson, Clare E. Mackay, and Masud Husain. 2019. Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical* 23 (Jan. 2019), 101904. https://doi.org/10.1016/j.nicl.2019.101904 Publisher: Elsevier Inc.

Gabor Perlaki, Gergely Orsi, Eniko Plozer, Anna Altbacker, Gergely Darnai, Szilvia Anett Nagy, Reka Horvath, Arnold Toth, Tamas Doczi, Norbert Kovacs, Peter Bogner, Attila Schwarcz, and Jozsef Janszky. 2014. Are there any gender differences in the hippocampus volume after head-size correction? A volumetric and voxel-based morphometric study. *Neuroscience Letters* 570 (June 2014), 119–123. https://doi.org/10.1016/j.neulet.2014.04.013 Publisher: Elsevier Ireland Ltd.

Peter S. Pressman, Yuliana Noniyeva, Nick Bott, Shubir Dutt, Virginia Sturm, Bruce L. Miller, and Joel H. Kramer. 2016. Comparing Volume Loss in Neuroanatomical Regions of Emotion versus Regions of Cognition in Healthy Aging. *PLOS ONE* 11, 8 (Aug. 2016), e0158187. https://doi.org/10.1371/journal.pone.0158187 Publisher: Public Library of Science.

Martin Reuter, H. Diana Rosas, and Bruce Fischl. 2010. Highly accurate inverse consistent registration: A robust approach. *NeuroImage* 53, 4 (Dec. 2010), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020

Martin Reuter, Nicholas J. Schmansky, H. Diana Rosas, and Bruce Fischl. 2012. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage* 61, 4 (July 2012), 1402–1418. https://doi.org/10.1016/j.neuroimage.2012.02.084

Dheeraj S. Roy, Takashi Kitamura, Teruhiro Okuyama, Sachie K. Ogawa, Chen Sun, Yuichi Obata, Atsushi Yoshiki, and Susumu Tonegawa. 2017. Distinct Neural Circuits for the Formation and Retrieval of Episodic Memories. *Cell* 170, 5 (Aug. 2017), 1000–1012.e19. https://doi.org/10.1016/j.cell.2017.07.013 Publisher: Cell Press.

Z. M. Saygin, D. Kliemann, J. E. Iglesias, A. J.W. van der Kouwe, E. Boyd, M. Reuter, A. Stevens, K. Van Leemput, A. McKee, M. P. Frosch, B. Fischl, and J. C. Augustinack. 2017. High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. *NeuroImage* 155 (July 2017), 370–382. https://doi.org/10.1016/j.neuroimage.2017.04.046 Publisher: Academic Press Inc.

Dorothee Schoemaker, Claudia Buss, Kevin Head, Curt A. Sandman, Elysia P. Davis, M. Mallar Chakravarty, Serge Gauthier, and Jens C. Pruessner. 2016.

Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *NeuroImage* 129 (April 2016), 1–14. https://doi.org/10.1016/j.neuroimage.2016.01.038 Publisher: Academic Press Inc.

F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl. 2004. A hybrid approach to the skull stripping problem in MRI. *NeuroImage* 22, 3 (July 2004), 1060–1075. https://doi.org/10.1016/j.neuroimage.2004.03.032 Publisher: Academic Press.

Christopher D. Whelan, Derrek P. Hibar, Laura S. Van Velzen, Anthony S. Zannas, Tania Carrillo-Roa, Katie Zubiraca McMahon, Gautam Prasad, Sinéad Kelly, Joshua Faskowitz, Greig De Zubiracay, Juan E. Iglesias, Theo G.M. Van Erp, Thomas Frodl, Nicholas G. Martin, Margaret J. Wright, Neda Jahanshad, Lianne Schmaal, Philipp G. Sämann, and Paul M. Thompson. 2016. Heritability and reliability of automatically segmented human hippocampal formation subregions. *NeuroImage* 128 (March 2016), 125–137. https://doi.org/10.1016/j.neuroimage.2015.12.039 Publisher: Academic Press Inc.

J. Wonderlick, D. Ziegler, P. Hosseini-Varnamkhasti, J. Locascio, A. Bakkour, A. Van Der Kouwe, C. Triantafyllou, S. Corkin, and B. Dickerson. 2009. Reliability of Mri-Derived Cortical and Subcortical Morphometric Measures: Effects of Pulse Sequence, Voxel Geometry, and Parallel Imaging. *NeuroImage* 44, 4 (2009), 1324–1333. https://doi.org/10.1016/j.neuroimage.2008.10.037

Amanda Worker, Danai Dima, Anna Combes, William R. Crum, Johannes Streffer, Steven Einstein, Mitul A. Mehta, Gareth J. Barker, Steve C. R. Williams, and Owen O'daly. 2018. Test–retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and Alzheimer's disease populations. *Human Brain Mapping* 39, 4 (2018), 1743–1754. https://doi.org/10.1002/hbm.23948

Kaan Yucel, Margaret C. McKinnon, Valerie H. Taylor, Kathryn Macdonald, Martin Alda, L. Trevor Young, and Glenda M. MacQueen. 2007. Bilateral Hippocampal Volume Increases After Long-Term Lithium Treatment in Patients With Bipolar Disorder: a Longitudinal Mri Study. *Psychopharmacology* 195, 3 (2007), 357–367. https://doi.org/10.1007/s00213-007-0906-9

Fenglian Zheng, Chuntong Li, Dongsheng Zhang, Dong Cui, Zhipeng Wang, and Jianfeng Qiu. 2019. Study on the sub-regions volume of hippocampus and amygdala in schizophrenia. *Quantitative Imaging in Medicine and Surgery* 9, 6 (2019), 1025–1036. https://doi.org/10.21037/qims.2019.05.21 Publisher: AME Publishing Company.

Xi Nian Zuo, Jeffrey S. Anderson, Pierre Bellec, Rasmus M. Birn, Bharat B. Biswal, Janusch Blautzik, John C.S. Breitner, Randy L. Buckner, Vince D.

Calhoun, F. Xavier Castellanos, Antao Chen, Bing Chen, Jiangtao Chen, Xu Chen, Stanley J. Colcombe, William Courtney, R. Cameron Craddock, Adriana Di Martino, Hao Ming Dong, Xiaolan Fu, Qiyong Gong, Krzysztof J. Gorgolewski, Ying Han, Ye He, Yong He, Erica Ho, Avram Holmes, Xiao Hui Hou, Jeremy Huckins, Tianzi Jiang, Yi Jiang, William Kelley, Clare Kelly, Margaret King, Stephen M. LaConte, Janet E. Lainhart, Xu Lei, Hui Jie Li, Kaiming Li, Kuncheng Li, Qixiang Lin, Dongqiang Liu, Jia Liu, Xun Liu, Yijun Liu, Guangming Lu, Jie Lu, Beatriz Luna, Jing Luo, Daniel Lurie, Ying Mao, Daniel S. Margulies, Andrew R. Mayer, Thomas Meindl, Mary E. Meyerand, Weizhi Nan, Jared A. Nielsen, David O'Connor, David Paulsen, Vivek Prabhakaran, Zhigang Qi, Jiang Qiu, Chunhong Shao, Zarrar Shehzad, Weijun Tang, Arno Villringer, Huiling Wang, Kai Wang, Dongtao Wei, Gao Xia Wei, Xu Chu Weng, Xuehai Wu, Ting Xu, Ning Yang, Zhi Yang, Yu Feng Zang, Lei Zhang, Qinglin Zhang, Zhe Zhang, Zhiqiang Zhang, Ke Zhao, Zonglei Zhen, Yuan Zhou, Xing Ting Zhu, and Michael P. Milham. 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data* 1 (Dec. 2014). `https://doi.org/10.1038/sdata.2014.49` Publisher: Nature Publishing Groups.