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

Nicholas J Buser[1], Christopher R. Madan[2], & Jamie L. Hanson[1]

1. University of Pittsburgh, Pittsburgh, PA, USA
2. University of Nottingham, Nottingham, United Kingdom

Corresponding Author:

Jamie Hanson, Ph.D.

Learning Research & Development Center

University of Pittsburgh

3939 O'Hara Street

Pittsburgh, PA 15260

Telephone: 412-383-3250

Email: jamie.hanson@pitt.edu

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 (*v7.1.0*). In a sample of participants with repeated structural imaging scans (N=118), we found numerical reliability (as assessed by intraclass correlations, ICC) was reasonable, with an average ICCs of 0.853 for hippocampal subfields, and 0.878 for amygdala subnuclei. However, only 26% of all subfields and subnuclei were "excellent" in terms of having numerical reliability $\geq$ 0.90. Spatial reliability was similarly reasonable, with 39% of hippocampal subfields and 33% of amygdala subnuclei having Dice coefficients $\geq$ 0.70. However, and of note, multiple regions (e.g., portions of the Dentate Gyrus; Cornu Ammonis 3; Medial and Paralaminar amygdala nuclei) had poor numerical and/or spatial reliability. We also examined correlations between spatial reliability and person-level factors (e.g., participant age and sex; inter-scan interval, and difference in image quality). For these factors, inter-scan interval and sex were related to variations in spatial reliability metrics. Examined collectively, our work suggests caution should be exercised for a few hippocampal subfields and amygdala nuclei with more variable reliability.

Keywords*: Amygdala; Hippocampus; Automated Segmentation; FreeSurfer*

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, 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 (Caldwell et al., 2015; Gunten et al., 2000; Gunten & Ron, 2004; Hanson et al., 2015; MacQueen et al., 2003; Malykhin et al., 2008; 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 (Madan & Kensinger, 2017). 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 & 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 work has 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 areas 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 & 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, 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 regard 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 numerical reliability and slightly lower spatial reliability. Numerical 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 numerical 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 numerical 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 pathology 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 numerical 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., 2020; Madan & 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

Participants

Data from an open-access neuroimaging initiative, the Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 2014), was used to investigate numerical 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.

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. MRI Images were visually inspected to determine if a participant's scans should be included in subsequent processing steps (e.g., FreeSurfer).

Structural Neuroimaging Processing (FreeSurfer)

Standard-processing approaches from FreeSurfer (e.g., cortical reconstruction; volumetric segmentation) were performed in version 7.1 (Stable Release, May 11, 2020) This was implemented via Brainlife.io (brainlife.app.0), a free, publicly funded, cloud-computing platform that allows for developing reproducible neuroimaging processing pipelines and sharing data (Pestilli 2018; Avesani et al. 2019). FreeSurfer is a widely-documented and freely available morphometric processing tool suite (http://surfer.nmr.mgh.harvard.edu). The technical details of FreeSurfer are described in prior publications (Dale et al., 1999; Fischl, 2012; Fischl et al., 2002, 2004; Fischl, Sereno, & Dale, 1999; Fischl, Sereno, Tootell, et al., 1999). 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.

FreeSurfer version 7.1 natively includes options to segment hippcampal subfields and

amygdalar nuclei. 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 12 subfields: (1) Parasubiculum, (2) Presubiculum [Head and Body], (3) Subiculum [Head and Body], (4) CA1 [Head and Body], (5) CA3 [Head and Body], (6) CA4 [Head and Body], (7) Granule Cell and Molecular Layer of the Dentate Gyrus [GC-ML-DG, Head and Body], (8) Molecular layer [Head and Body], (9) Fimbria, (10) Hippocampal Fissure, (11) Hippocampal Tail, and (12) Hippocampus-Amygdala-Transition-Area (HATA). This yields nineteen subdivisions from FreeSurfer (including these regions and head/body divisions).

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 6 subnuclei: (1) Lateral, (2) Basal, (3) Central, (4) Medial, (5) Cortical, (6) Accessory Basal, (6) Paralaminar. Two additional

subdivisions, the Corticoamygdaloid Transition Area and Anterior Amygdaloid Area, are also output.

Of note, here we processed two scans from each participant using the “cross-sectional” pipeline. This is in contrast to FreeSurfer's longitudinal stream that creates an unbiased within-subject template image to improve registration (Reuter et al., 2010; Reuter et al., 2012). Cross-sectional pipelines were applied to both scans (test and retest) for each participant. For both the hippocampal subfields and amygdala nuclei, volume (in mm^3) for each subdivision was extracted and used in numerical reliability analysis. Spatial information (labelled voxels in axial, coronal, and spatial orientations) was also output for each subdivision, for use in spatial reliability analysis. Of note, we have also implemented FreeSurfer's longitudinal processing stream and report on these results in our Supplemental Materials.

Automated MRI Image 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 & 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 averaged the score for each participant's first scan from their second, using that as a

measure of interscan quality difference.

Derivation of Reliability Measures

To assess the reliability of the numerical 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 numerical cohesion across the given groups (McGraw & 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 & Lemon, 2012). A two-way model with absolute agreement was used in order to investigate the reliability of subdivision segmentation; this was calculated for each subdivision's volume (in mm^3). 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 & Li, 2016). In addition to ICCs, Bland-Altman metrics were calculated for each hippocampal and amygdala subdivision. In this approach, the mean difference ("*bias*") between the FreeSurfer outputs (from the first and second scan) was calculated and then presented as a proportion of the mean volume (Bland and Altman, 1986; Myles & Cui, 2007). Bland-Altman plots were also constructed for a few subdivisions to assess agreement between FreeSurfer outputs.

Although ICCs and Bland-Altman metrics serve as indicators of numerical reliability, these may still be incomplete, particularly when we think about the spatial information present in MRI volumes. Indeed, even with numerical similarity, there may be discrepancies in the specific spatial voxels labelled 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 (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. As recommended by past reports (Zijdenbos et al., 1994; Zou et al., 2004), we considered Dice Coefficients $\geq$0.700 as exhibiting “good” spatial overlap.

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., 2020). Noisier

images 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 - Scan 1) was included in these analyses.

3. Results

Hippocampus Reliability

Using ICC analysis, we found consistently reasonable levels of numerical reliability for hippocampal subfields. Multiple regions demonstrated “excellent” reliability (ICC $\geq$ 0.90), while the preponderance of subfields were in the “good” range (ICC=0.75-0.90). See Table 1 for values from the 19 subfield segmentations in each hemisphere. Of note, the head of CA3 in the right hemisphere demonstrated poor spatial reliability (0.737). Bland-Altman “bias” also showed a great deal of variability with differences between scans, as a portion of that structure’s volume, ranges from 0.237-5.639%. Using Dice coefficients as metrics of spatial reliability, results became a bit more variable with 5 areas showing poor spatial reliability and Dice Coefficients < 0.5. 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). Reliability metrics were generally higher using the FreeSurfer longitudinal processing stream; this data is detailed in our Supplemental Materials.

---- Table 1 Approximately Here ---

---- Figure 1 Approximately Here ---

Amygdala Reliability

Within the amygdala, the numerical reliability was mostly high (66% of subnuclei > 0.88). The Medial Nucleus, in both the left and right hemisphere, was the lowest in numerical reliability (Left: 0.794; Right: 0.791). Values are reported (*see Table 2*). Regarding spatial reliability, a few areas demonstrated good reliability (≥.8) including the lateral, basal, and accessory basal subnuclei. There were, however, areas with poor spatial reliability (Dice Coefficients=0.33-0.4). These problems were primarily seen in the Medial and Paralaminar Nuclei. This poor spatial reliability is graphically depicted in Figure 3, with the Paralaminar nucleus segments noted for the first and second scan from an exemplar subject. In contrast, a region, the Lateral amygdala nucleus, with good spatial reliability is shown in Figure 4. 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). Reliability metrics were, again, generally higher using the FreeSurfer longitudinal processing stream; this data is detailed in our Supplemental Materials.

---- Figure 2 Approximately Here ---

---- Table 2 Approximately Here ---

---- Figures 3 and 4 Approximately Here ---

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

We next examined associations between spatial reliability and subject-level variables. Correlations between the Hippocampal-subfield Dice coefficients and our subject-level

variables are shown in Figures 5-8. Length between neuroimaging sessions (test and retest) had a significant negative correlation ($p<.05$) with dice-coefficients for approximately 8 regions (shown in Figure 5). Differences in image quality were related to volumes in 2 subfields (also shown in Figure 5). Sex was related to the dice coefficients for approximately 12 subfields, while age was correlated with 2 subfield volumes.

---- *Table 3 Approximately Here* ---

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

---- *Table 4 Approximately Here* ---

4. Discussion

In this paper, we assessed the numerical and spatial reliability of FreeSurfer's hippocampal and amygdala sub-segmentation algorithms. The ICC's, serving as our indicator of numerical reliability, were reasonable (hippocampal subfields: 0.737-0.956; amygdala nuclei: 0.794-0.952), indicating that FreeSurfer is generally numerically reliable in providing overall volume for each subregion. Using Bland–Altman metrics of bias as an additional proxy of numerical reliability suggests a few regions exhibited a good bit of variability in segmentation from one scan to the next; specifically, 8 regions showed 2% in volume from one scan to the next. This is perhaps concerning given that individuals with

serious pathologies (i.e., Alzheimer's disease) often only differ 2-5% from control groups in subcortical volumes (e.g., Jack et al., 2005). The dice coefficients, serving as our indicator of spatial reliability, were reasonable, though slightly lower than the ICCs. Of potential concern, a few subdivisions, in both the hippocampus and amygdala, had fairly low spatial reliability, suggesting unreliable segmentation. Examined collectively, applied researchers should take care when applying these types of automated segmentation techniques, especially if not thoroughly trained in amygdala and hippocampal anatomy.

While our results suggest that many of the volumetric outputs of amygdala and hippocampal subdivisions are mostly numerically reliable, the drop in spatial reliability may mean researchers should exercise caution in the analysis and interpretation of areas with poor spatial reliability. For example, the hippocampal fissure, paralaminar nucleus (amygdala), medial nucleus (amygdala) showed poor spatial reliability ($\leq$.4) through their Dice coefficients. Because the spatial reliability of these areas is relatively poor, studies that interpret changes in volume within or across subjects might be using segmentations which contain improperly (or inconsistently) classified voxels within those regions. For example, several studies have already reported significant findings from the paralaminar nucleus of the amygdala (Morey et al., 2020; Zheng et al., 2019); given the questionable reproducibility of its anatomical bounding, these findings may require further verification.

Connected to spatial reliability, there are a few key drivers of the substandard performance in this domain. First, these areas are fairly small in size and may be difficult to isolate. In such cases, even a few mislabelled voxels can greatly influence spatial overlap. Many of the areas with the lowest spatial reliability are also the smallest subdivisions. For example, the Paralaminar and the medial nuclei of the amygdala range between 20-60 mm^3 in our

sample and have some of the lowest spatial reliability values. However, and of note, this is not the only factor hampering performance, as other structures (of similar sizes) have reasonable spatial reliability values (e.g., HATA ≥ 0.643 Dice Coefficients; Parasubiculum ≥ 0.627 Dice Coefficients), while comparatively larger structures (e.g., hippocampal-fissure; CA3-body) demonstrate lower spatial reliability. Second, irregular MR contrast is often common to these areas, especially for the amygdala. Given the close vicinity to bone and sinuses, there is typically susceptibility-related dropout, field inhomogeneities and physiological artifacts in the amygdala and the hippocampus (Merboldt et al., 2001; Robinson et al., 2004; Windischberger et al., 2002). This may introduce inconsistent gray/white matter contrast, complicating isolation of different subdivisions. Finally, a number of amygdala and hippocampal subdivisions are irregularly and complexly shaped. For example, both the anterior and posterior borders of the amygdala are difficult to consistently demarcate (Achten et al., 1998; Convit et al., 1999; Watson et al., 1992, as cited in Entis et al., 2012). Many past reports employing manual training actually employ “heuristics” rather than clear anatomical boundaries (e.g., Nacewicz et al., 2006 used a “semicircle substitution”).

Given these challenges, it will be important to think about novel approaches to break down the amygdala and hippocampus into smaller subdivisions, while still maintaining high validity and reproducibility. In regard to the hippocampus, there has been a great deal of progress made by the Hippocampal Subfields Group (https://www.hippocampalsubfields.com/); this is a collaborative of >200 imaging and anatomy experts worldwide that has established guidelines for appropriate MRI acquisition for researchers interested in the hippocampus, as well as developing candidate protocols for segmentation of hippocampal subregions. This group has published protocol

comparison and harmonization papers (Wisse et al., 2017, Yushkevich et al., 2015), as well as important commentaries on the issue (e.g., Olsen et al., 2019; Wisse et al., 2020). Similar joint efforts are not, to our knowledge, currently underway for amygdala subnuclei segmentation. Convening such a collaborative could be particularly powerful moving forward, especially as debate has been fairly continuous regarding subdivisions of the amygdala at the histological level (e.g., Swanson and Petrovich, 1998). In the interim, or in the absence of that type of group, it may be reasonable to only consider more macro-level amygdala segmentation (e.g., basolateral, centromedial, basomedial, and amygdaloid cortical complexes, as detailed by Mai, Majtanik, & Paxinos, 2015). Many groups have moved towards this idea, aggregating subdivisions using latent factor modelling and other techniques to group related regions (e.g., Oshri et al., 2019).

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 should be pursued to establish the validity of these FreeSurfer subdivisions. 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; Aghamohammadi-Sereshki et al., 2018); however, this typically requires high-resolution and high-field strength neuroimaging (i.e., >3T MRI Scanner, sub-millimeter voxels). Studies looking at the degree of overlap between such methods and the FreeSurfer algorithm for amygdala segmentation would be helpful for effective evaluation of validity.

Moving forward, our results suggest a number of important future directions, ranging from

study recommendations to quantitative directions. In keeping with a recent commentary by Wisse et al. (2020), we might, at present, caution against wide-spread usage of volumetric measurements of hippocampal subfields and amygdala subdivisions especially if there are to be derived from “standard” T1-weighted sequences (e.g., 1x1x1 mm^3 resolution). As noted by Wisse and colleagues, the internal structures of these subcortical subdivisions can rarely be visualized at “standard” resolutions. Instead, we would encourage researchers interested in these subdivisions to use higher resolution anatomical (T1-weighted) scans, or at minimum use both T1- and T2-weighted scans. While this is a strong statement, many studies focused on hippocampal subfield volumes have reported inconsistent and contradictory (and many would suggest “biologically implausible”) results (Haukvik, Tamnes, Söderman, & Agartz, 2018; Wisse et al., 2020). We worry this trend will soon be seen for amygdala subnuclei. At the bare minimum, we would encourage research groups to only examine larger subdivisions with reasonable numerical and spatial reliability (e.g., Lateral, Basal, and Accessory Basal Nuclei for the Amygdala; Head, Body, and Tail for the Hippocampus). Connected to ideas of multimodal neuroimaging, we believe deployment of novel deep-learning approaches (e.g., Dilated-Dense U-Net, Li et al., 2019; Cycle-consistent generative adversarial networks, Liu et al., 2020) may aid in dealing with the small size, varying MRI contrast, and complex shapes of different subdivisions. We believe deployment of such methods using T1- and T2-weighted scans, as well as diffusion weighted imaging, from the same participants could produce more reliable amygdala and hippocampal segments.

In considering our results, it is important to note a few potential limitations of our work. First, in our main manuscript, we processed repeated MRI images using the cross-sectional FreeSurfer pipeline (and not the longitudinal stream of this software suite). This

may be a debated choice. However, FreeSurfer's longitudinal pipeline is not actually independently "considering" (segmenting) the different MRI scans. This may violate some theoretical aspects of test-retest reliability and produce a more favorable set of reliability metrics for FreeSurfer's methods. To allow complete consideration of this idea, we have included results from the longitudinal pipeline in our Supplemental Materials. Second, the sample is a rather homogenous group of individuals and may not represent the greater population. All participants were recruited from Beijing Normal University, with limited race and ethnic variability. Similarly, all participants were neurotypical young adults in a fairly constrained age range (Mean=22.08+/-1.88). Due to these demographics, it is uncertain to what degree these findings will generalize outside of that domain and further research should be done to ascertain this. With past work noting brain differences between samples from China and the United States of America (Shi et al., 2017; Xie et al., 2015), 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. Other studies examining hippocampal subfield segmentation have used diverse sets of participants, so this may be less of an issue for this structure. Additional work focusing on the amygdala subnuclei using cohorts with greater variability, clinical status, and other sociodemographic factors should, however, be pursued to overcome this limitation. Third, we do not have detailed information regarding more fine-grain drivers of differences in image quality. We examined an aggregated measure of image quality but did not see consistent associations with segmentations of subdivisions. This was slightly surprising to our group. 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.

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 in FreeSurfer 7. 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.

Information Sharing Statement: Neuroimaging data used in our analyses were sourced from the Consortium for Reliability and Reproducibility Study (CoRR, Zuo et al. 2014). FreeSurfer software is publicly and freely available from the FreeSurferWiki resource (http://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki), which is developed and maintained at the Martinos Center for Biomedical Imaging (http://www.nmr.mgh.harvard.edu/martinos/noFlashHome.php). This software, information and support are provided online at the FreeSurferWiki webpage.

Funding: This work was support by internal funds provided by the University of Pittsburgh. In addition, computing hardware and software was also provided through the University og Pittsburgh's Center for Research Computing (Pitt CRC) (www.crc.pitt.edu).

Conflicts of interest/Competing interests: None

References

Achten, E., Deblaere, K., De Wagter, C., Van Damme, F., Boon, P., De Reuck, J., & Kunnen, M. (1998). Intra-and interobserver variability of MRI-based volume measurements of the hippocampus and amygdala using the manual ray-tracing method. Neuroradiology, 40(9), 558-566.

Aghamohammadi-Sereshki, A., Huang, Y., Olsen, F., & Malykhin, N. V. (2018). In vivo quantification of amygdala subnuclei using 4.7 T fast spin echo imaging. Neuroimage, 170, 151-163.

Avesani, P., McPherson, B., Hayashi, S., Caiafa, C.F., Henschel, R., Garyfallidis, E., Kitchell, L., Bullock, D., Patterson, A., Olivetti, E., Sporns, O., Saykin, A.J., Wang, L., Dinov, I., Hancock, D., Caron, B., Qian, Y., Pestilli, F., 2019. The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. Scientific Data 6, 69. https://doi.org/ 10.1038/s41597-019-0073-y

Bittner, K. C., Grienberger, C., Vaidya, S. P., Milstein, A. D., Macklin, J. J., Suh, J., Tonegawa, S., & Magee, J. C. (2015). Conjunctive input processing drives feature selectivity in hippocampal CA1 neurons. *Nature Neuroscience*, *18*(8), 1133–1142. https://doi.org/10.1038/nn.4062

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307-310.

Brown, E. M., Pierce, M. E., Clark, D. C., Fischl, B. R., Iglesias, J. E., Milberg, W. P., McGlinchey, R. E., & Salat, D. H. (2020). Test-retest reliability of FreeSurfer automated hippocampal subfield segmentation within and across scanners. *NeuroImage*, *210*. https://doi.org/10.1016/j.neuroimage.2020.116563

Caldwell, J. Z., Armstrong, J. M., Hanson, J. L., Sutterer, M. J., Stodola, D. E., Koenigs,

M., ... & Davidson, R. J. (2015). Preschool externalizing behavior predicts gender-specific variation in adolescent neural structure. PloS one, 10(2), e0117453.

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

Convit, A., McHugh, P., Wolf, O. T., De Leon, M. J., Bobinski, M., De Santi, S., ... & Tsui, W. (1999). MRI volume of the amygdala: a reliable method allowing separation from the hippocampal formation. Psychiatry Research: Neuroimaging, 90(2), 113-123.

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

Dahnke, R., Ziegler, G., Grosskreutz, J., & Gaser, C. (2015). Quality Assurance in Structural MRI. *21st Annual Meeting of the Organization For Human Brain Mapping*, 1556.

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

Daugherty, A. M., Bender, A. R., Raz, N., & Ofen, N. (2016). Age differences in hippocampal subfield volumes from childhood to late adulthood. *Hippocampus*, *26*(2), 220–228. https://doi.org/10.1002/hipo.22517

Entis, J. J., Doerga, P., Barrett, L. F., & Dickerson, B. C. (2012). A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution MRI. *NeuroImage*, *60*(2), 1226–1235. https://doi.org/10.1016/j.neuroimage.2011.12.073

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781.

https://doi.org/10.1016/j.neuroimage.2012.01.021

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Fischl, B., Salat, D. H., Van Der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23*(SUPPL. 1). https://doi.org/10.1016/j.neuroimage.2004.07.016

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*(2), 195–207. https://doi.org/10.1006/nimg.1998.0396

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

Game, M., Lemon J., Fellows I., & Singh, P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

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

Gilmore, A., Buser, N., & Hanson, J. L. (2020). Variations in Structural MRI Quality Significantly Impact Commonly-Used Measures of Brain Anatomy. *biorxiv*, 581876.

https://doi.org/10.1101/581876

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

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

Guzman, S. J., Schlögl, A., Frotscher, M., & Jonas, P. (2016). Synaptic mechanisms of pattern completion in the hippocampal CA3 network. *Science*, *353*(6304), 1117–1123. https://doi.org/10.1126/science.aaf1836

Hanson, J. L., Nacewicz, B. M., Sutterer, M. J., Cayo, A. A., Schaefer, S. M., Rudolph, K. D., ... & Davidson, R. J. (2015). Behavioral problems after early life stress: contributions of the hippocampus and amygdala. Biological psychiatry, 77(4), 314-323.

Hanson, J. L., Suh, J. W., Nacewicz, B. M., Sutterer, M. J., Cayo, A. A., Stodola, D. E., Burghy, C. A., Wang, H., Avants, B. B., Yushkevich, P. A., Essex, M. J., Pollak, S. D., & Davidson, R. J. (2012). Robust Automated Amygdala Segmentation via Multi-Atlas Diffeomorphic Registration. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00166

Haukvik, U. K., Tamnes, C. K., Söderman, E., & Agartz, I. (2018). Neuroimaging hippocampal subfields in schizophrenia and bipolar disorder: a systematic review and meta-analysis. Journal of Psychiatric Research, 104, 217-226.

Herten, A., Konrad, K., Krinzinger, H., Seitz, J., & von Polier, G. G. (2019). Accuracy and bias of automatic hippocampal segmentation in children and adolescents. *Brain*

Structure and Function, *224*(2), 795–810. https://doi.org/10.1007/s00429-018-1802-2

Huang, L., Huang, T., Zhen, Z., & Liu, J. (2016). A test-retest dataset for assessing long-Term reliability of brain morphology and resting-state brain activity. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.16

Iglesias, J. E., Augustinack, J. C., Nguyen, K., Player, C. M., Player, A., Wright, M., Roy, N., Frosch, M. P., McKee, A. C., Wald, L. L., Fischl, B., & 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., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., Reuter, M, & Alzheimer's Disease Neuroimaging. (2016). *Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases*. Neuroimage, 141, 542-555.

Jack, C. R., Shiung, M. M., Weigand, S. D., O'brien, P. C., Gunter, J. L., Boeve, B. F., ... & Petersen, R. C. (2005). Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. Neurology, 65(8), 1227-1231.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., … and, M. W. W. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): Mri Methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049

Jovicich, J., Czanner, S., Han, X., Salat, D., Kouwe, A. van der, Quinn, B., Pacheco, J., Albert, M., Killiany, R., & Blacker, D. (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), 177–192. https://doi.org/10.1016/j.neuroimage.2009.02.010

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F., Hensch, T., Tränkner, A., Schönknecht, P., Leroy, M., Lopes, R., Bordet, R., Chanoine, V., Ranjeva, J.-P., Didic, M., … Frisoni, G. B. (2013). Brain Morphometry Reproducibility in Multi-Center 3t Mri Studies: A Comparison of Cross-Sectional and Longitudinal Segmentations. *NeuroImage*, *83*(nil), 472–484. https://doi.org/10.1016/j.neuroimage.2013.05.007

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

Liem, F., Mérillat, S., Bezzola, L., Hirsiger, S., Philipp, M., Madhyastha, T., & Jäncke, L. (2015). Reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly. *NeuroImage*, *108*, 95–109. https://doi.org/10.1016/j.neuroimage.2014.12.035

Li, G., Chen, M. H., Li, G., Wu, D., Lian, C., Sun, Q., ... & Wang, L. (2019, October). A Longitudinal MRI Study of Amygdala and Hippocampal Subfields for Infants with Risk of Autism. In International Workshop on Graph Learning in Medical Imaging (pp. 164-171). Springer, Cham.

Lin, Q., Dai, Z., Xia, M., Han, Z., Huang, R., Gong, G., Liu, C., Bi, Y., & He, Y. (2015). A connectivity-based test-retest dataset of multi-modal magnetic resonance imaging in young healthy adults. *Scientific Data*, *2*. https://doi.org/10.1038/sdata.2015.56

Liu, Y., Nacewicz, B. M., Zhao, G., Adluru, N., Kirk, G. R., Ferrazzano, P. A., ... & Alexander, A. L. (2020). A 3D Fully Convolutional Neural Network With Top-Down Attention-Guided Refinement for Accurate and Robust Automatic Segmentation of Amygdala and Its Subnuclei. Frontiers in Neuroscience, 14, 260.

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

Madan, C. R., & Kensinger, E. A. (2017). Test–retest reliability of brain morphology estimates. *Brain Informatics*, *4*(2), 107–121. https://doi.org/10.1007/s40708-016-0060-4

Mai, J. K., Majtanik, M., & Paxinos, G. (2015). Atlas of the human brain. Academic Press.

Malykhin, N. V., Bouchard, T. P., Camicioli, R., & Coupland, N. J. (2008). Aging hippocampus and amygdala. *NeuroReport*, *19*(5), 543–547. https://doi.org/10.1097/WNR.0b013e3282f8b18c

Martino, A. D., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2013). The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Marwha, D., Halari, M., & Eliot, L. (2017). Meta-analysis reveals a lack of sexual dimorphism in human amygdala volume. *NeuroImage*, *147*, 282–294. https://doi.org/10.1016/j.neuroimage.2016.12.021

McGraw, K. O., & Wong, S. P. (1996). Forming Inferences about Some Intraclass

Correlation Coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Merboldt, K. D., Fransson, P., Bruhn, H., & Frahm, J. (2001). Functional MRI of the human amygdala?. Neuroimage, 14(2), 253-257.

Morey, R. A., Clarke, E. K., Haswell, C. C., Phillips, R. D., Clausen, A. N., Mufford, M. S., Saygin, Z., Brancu, M., Beckham, J. C., Calhoun, P. S., Dedert, E., Elbogen, E. B., Fairbank, J. A., Hurley, R. A., Kilts, J. D., Kimbrel, N. A., Kirby, A., Marx, C. E., McDonald, S. D., … LaBar, K. S. (2020). Amygdala Nuclei Volume and Shape in Military Veterans With Posttraumatic Stress Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(3), 281–290. https://doi.org/10.1016/j.bpsc.2019.11.016

Morey, R. A., Petty, C. M., Xu, Y., Pannu Hayes, J., Wagner, H. R., Lewis, D. V., LaBar, K. S., Styner, M., & McCarthy, G. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *NeuroImage*, *45*(3), 855–866. https://doi.org/10.1016/j.neuroimage.2008.12.033

Morey, R. A., Selgrade, E. S., Wagner, H. R., Huettel, S. A., Wang, L., & McCarthy, G. (2010). Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Human Brain Mapping*, *31*(11), 1751–1762. https://doi.org/10.1002/hbm.20973

Myles, P. S., & Cui, J. (2007). I. Using the Bland–Altman method to measure agreement with repeated measures. British Journal of Anaesthesia, 99(3), 309–311, https://doi.org/10.1093/bja/aem214

Nacewicz, B. M., Dalton, K. M., Johnstone, T., Long, M. T., McAuliff, E. M., Oakes, T. R., ... & Davidson, R. J. (2006). Amygdala volume and nonverbal social impairment in adolescent and adult males with autism. Archives of general psychiatry, 63(12),

1417-1428.

Neunuebel, J. P., & Knierim, J. J. (2014). CA3 retrieves coherent representations from degraded input: Direct evidence for CA3 pattern completion and dentate gyrus pattern separation. *Neuron*, *81*(2), 416–427. https://doi.org/10.1016/j.neuron.2013.11.017

Nobis, L., Manohar, S. G., Smith, S. M., Alfaro-Almagro, F., Jenkinson, M., Mackay, C. E., & Husain, M. (2019). Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical*, *23*, 101904. https://doi.org/10.1016/j.nicl.2019.101904

Olsen, R. K., Carr, V. A., Daugherty, A. M., La Joie, R., Amaral, R. S., Amunts, K., ... & Boccardi, M. (2019). Progress update from the hippocampal subfields group. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 11(1), 439-449.

Oshri, A., Gray, J. C., Owens, M. M., Liu, S., Duprey, E. B., Sweet, L. H., & MacKillop, J. (2019). Adverse childhood experiences and amygdalar reduction: High-resolution segmentation reveals associations with subnuclei and psychiatric outcomes. Child maltreatment, 24(4), 400-410.

Perlaki, G., Orsi, G., Plozer, E., Altbacker, A., Darnai, G., Nagy, S. A., Horvath, R., Toth, A., Doczi, T., Kovacs, N., Bogner, P., Schwarcz, A., & Janszky, J. (2014). Are there any gender differences in the hippocampus volume after head-size correction? A volumetric and voxel-based morphometric study. *Neuroscience Letters*, *570*, 119–123. https://doi.org/10.1016/j.neulet.2014.04.013

Pestilli, F., 2018. Human white matter and knowledge representation. PLoS Biol. 16, e2005758. https://doi.org/ 10.1371/journal.pbio.2005758

Pressman, P. S., Noniyeva, Y., Bott, N., Dutt, S., Sturm, V., Miller, B. L., & Kramer, J. H.

(2016). Comparing Volume Loss in Neuroanatomical Regions of Emotion versus Regions of Cognition in Healthy Aging. *PLOS ONE*, *11*(8), e0158187. https://doi.org/10.1371/journal.pone.0158187

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, *53*(4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020

Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, *61*(4), 1402–1418. https://doi.org/10.1016/j.neuroimage.2012.02.084

Robinson, S., Windischberger, C., Rauscher, A., & Moser, E. (2004). Optimized 3 T EPI of the amygdalae. Neuroimage, 22(1), 203-210.

Roy, D. S., Kitamura, T., Okuyama, T., Ogawa, S. K., Sun, C., Obata, Y., Yoshiki, A., & Tonegawa, S. (2017). Distinct Neural Circuits for the Formation and Retrieval of Episodic Memories. *Cell*, *170*(5), 1000–1012.e19. https://doi.org/10.1016/j.cell.2017.07.013

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

Schoemaker, D., Buss, C., Head, K., Sandman, C. A., Davis, E. P., Chakravarty, M. M., Gauthier, S., & Pruessner, J. C. (2016). Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *NeuroImage*, *129*, 1–14. https://doi.org/10.1016/j.neuroimage.2016.01.038

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

Shi, L., Liang, P., Luo, Y., Liu, K., Mok, V. C., Chu, W. C., ... & Li, K. (2017). Using large-scale statistical Chinese brain template (Chinese2020) in popular neuroimage analysis toolkits. Frontiers in human neuroscience, 11, 414.

Swanson, L. W., & Petrovich, G. D. (1998). What is the amygdala?. Trends in neurosciences, 21(8), 323-331.

Watson, C., Andermann, F., Gloor, P., Jones-Gotman, M., Peters, T., Evans, A., ... & Leroux, G. (1992). Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. Neurology, 42(9), 1743-1743.

Whelan, C. D., Hibar, D. P., Van Velzen, L. S., Zannas, A. S., Carrillo-Roa, T., McMahon, K. Z., Prasad, G., Kelly, S., Faskowitz, J., De Zubiracay, G., Iglesias, J. E., Van Erp, T. G. M., Frodl, T., Martin, N. G., Wright, M. J., Jahanshad, N., Schmaal, L., Sämann, P. G., & Thompson, P. M. (2016). Heritability and reliability of automatically segmented human hippocampal formation subregions. *NeuroImage*, *128*, 125–137. https://doi.org/10.1016/j.neuroimage.2015.12.039

Windischberger, C., Langenberger, H., Sycha, T., Tschernko, E. M., Fuchsjäger-Mayerl, G., Schmetterer, L., & Moser, E. (2002). On the origin of respiratory artifacts in BOLD-EPI of the human brain. Magnetic resonance imaging, 20(8), 575-582.

Wisse, L. E., Chételat, G., Daugherty, A. M., de Flores, R., la Joie, R., Mueller, S. G., ... & Raz, N. (2020). Hippocampal subfield volumetry from structural isotropic 1 mm3 MRI scans: A note of caution. Human Brain Mapping.

Wisse, L. E., Daugherty, A. M., Olsen, R. K., Berron, D., Carr, V. A., Stark, C. E., ... & Bernstein, J. D. (2017). A harmonized segmentation protocol for hippocampal and

parahippocampal subregions: Why do we need one and what are the key goals?. Hippocampus, 27(1), 3-11.

Wonderlick, J. S., Ziegler, D. A., Hosseini-Varnamkhasti, P., Locascio, J. J., Bakkour, A., Van Der Kouwe, A., ... & Dickerson, B. C. (2009). Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. *Neuroimage*, *44*(4), 1324-1333. https://doi.org/10.1016/j.neuroimage.2008.10.037

Worker, A., Dima, D., Combes, A., Crum, W. R., Streffer, J., Einstein, S., Mehta, M. A., Barker, G. J., C. R. Williams, S., & O'daly, O. (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), 1743–1754. https://doi.org/10.1002/hbm.23948

Xie, W., Richards, J. E., Lei, D., Lee, K., & Gong, Q. (2015). Comparison of the brain development trajectory between Chinese and US children and adolescents. Frontiers in systems neuroscience, 8, 249.

Yucel, K., McKinnon, M. C., Taylor, V. H., Macdonald, K., Alda, M., Young, L. T., & MacQueen, G. M. (2007). Bilateral Hippocampal Volume Increases After Long-Term Lithium Treatment in Patients With Bipolar Disorder: A Longitudinal Mri Study. *Psychopharmacology*, *195*(3), 357–367. https://doi.org/10.1007/s00213-007-0906-9

Yushkevich, P. A., Amaral, R. S., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., ... & Chételat, G. (2015). Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: towards a harmonized segmentation protocol. Neuroimage, 111, 526-541.

Zheng, F., Li, C., Zhang, D., Cui, D., Wang, Z., & Qiu, J. (2019). Study on the sub-regions volume of hippocampus and amygdala in schizophrenia. *Quantitative Imaging in*

Medicine and Surgery, *9*(6), 1025–1036. https://doi.org/10.21037/qims.2019.05.21

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., & Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. IEEE transactions on medical imaging, 13(4), 716-724.

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., ... & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index: Scientific reports. Academic radiology, 11(2), 178-189.

Zuo, X. N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., Breitner, J. C. S., Buckner, R. L., Calhoun, V. D., Castellanos, F. X., Chen, A., Chen, B., Chen, J., Chen, X., Colcombe, S. J., Courtney, W., Craddock, R. C., Di Martino, A., Dong, H. M., … Milham, M. P. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, *1*. https://doi.org/10.1038/sdata.2014.49

Table 1.

Region	ICC LH	ICC RH	DICE LH	DICE RH	BIAS AS POV LH	BIAS AS POV RH
Parasubiculum	0.842	0.776	0.667	0.627	-3.729%	5.639%
Presubiculum Head	0.924	0.896	0.746	0.739	-1.282%	-1.513%
Presubiculum Body	0.909	0.868	0.743	0.725	1.854%	-3.363%
Subiculum Head	0.956	0.914	0.723	0.719	-0.857%	-0.237%
Subiculum Body	0.895	0.905	0.770	0.766	-2.185%	-1.038%
CA1 Head	0.899	0.931	0.752	0.757	-0.772%	-1.066%
CA1 Body	0.885	0.849	0.646	0.665	-1.214%	-0.130%
CA3 Head	0.8084	0.737	0.543	0.502	-1.876%	-1.764%
CA3 Body	0.882	0.789	0.506	0.494	-1.265%	-1.569%
CA4 Head	0.809	0.813	0.716	0.709	-1.190%	-1.296%
CA4 Body	0.779	0.779	0.714	0.691	-0.811%	-1.643%
GC ML DG Head	0.825	0.861	0.507	0.500	-1.080%	-1.069%
GC ML DG Body	0.759	0.777	0.496	0.468	-1.216%	-1.069%
Molecular Layer HP Head	0.897	0.926	0.605	0.604	-1.037%	0.830%
Molecular Layer	0.900	0.867	0.535	0.539	-1.375%	-0.996%

HP Body						
Fimbria	0.846	0.816	0.608	0.587	-1.506%	-2.774%
Hippocampal Fissure	0.769	0.861	0.368	0.417	1.259%	0.375%
Hippocampal Tail	0.946	0.879	0.845	0.832	1.261%	2.571%
HATA	0.867	0.786	0.661	0.643	-0.936%	-2.101%

Table 1: Intraclass Correlation Coefficients (ICC), Dice Coefficients (DICE), and Bland–Altman bias as a portion of a volume's structure (BIAS AS POV) for Hippocampal Subfields for left and right hemisphere regions (e.g., ICC LH = Intraclass Correlation Coefficients for left hemisphere; DICE RH = Dice Coefficient for right hemisphere). Color coding is in accordance with excellent [green], good [yellow], poor [red] scores for ICCs and Dice Coefficients (ICC: 0.90-1.00 [excellent], 0.75-0.90 [good], 0.00-0.75 [poor]; Dice Coefficients: 0.70-1.00 [excellent], 0.50-.70 [good], 0-0.50 [poor]). We have also highlighted regions with >2% bias as a portion of a volume's structure in yellow. Subfield Abbreviations include: Cornu Ammonis (CA), Granule Cell and Molecular Layer of Dentate Gyrus (GC-ML-DG); Hippocampus-Amygdala-Transition-Area (HATA); Hippocampal Parcellation (HP).

Table 2.

Region	ICC LH	ICC RH	DICE LH	DICE RH	BIAS AS POV LH	BIAS AS POV RH
Lateral Nucleus	0.952	0.943	0.876	0.868	-0.664%	-0.685%
Basal Nucleus	0.921	0.912	0.858	0.847	-0.216%	-1.04%
Central Nucleus	0.882	0.813	0.527	0.508	-1.841%	-1.329%
Medial Nucleus	0.794	0.791	0.339	0.308	-3.004%	-0.607%
Cortical Nucleus	0.867	0.812	0.511	0.521	-0.584%	0.477%
Accessory Basal Nucleus	0.934	0.913	0.848	0.833	-0.298%	-0.657%
Paralaminar Nucleus	0.890	0.821	0.394	0.383	-0.389%	-0.490%
Corticoamygdaloid Transition	0.898	0.881	0.698	0.684	-0.701%	-1.504%
Anterior Amygdaloid Area	0.897	0.892	0.597	0.583	-0.794%	1.491%

Table 2: Intraclass Correlation Coefficients (ICCs), Dice Coefficients (DICE), and Bland–Altman bias as a portion of a volume's structure (BIAS AS POV) for Amygdala Subnuclei for left and right hemisphere regions (e.g., ICC LH = Intraclass Correlation Coefficients for left hemisphere; DICE RH = Dice Coefficient for right hemisphere). Color coding is in accordance with excellent [green], good [yellow], poor [red] scores for ICCs and Dice Coefficients (ICC: 0.90-1.00 [excellent], 0.75-0.90 [good], 0.00-0.75 [poor]; Dice Coefficients: 0.70-1.00 [excellent], 0.50-.70 [good], 0-0.50 [poor]). We have also highlighted regions with >2% bias as a portion of a volume's structure in yellow.

Table 3.

Region	ISI r DICE		MRIQ r DICE		Sex r DICE		Age r DICE	
	LH	RH	LH	RH	LH	RH	LH	RH
Parasubiculum	-.10	-.05	.04	-.03	.13	.17	-.08	-.08
Presubiculum Head	-.15	-.04	.15	-.01	.23	.19	.06	-.03
Presubiculum Body	-.16	.03	.16	-.05	.16	.15	.10	-.02
Subiculum Head	-.16	.00	.05	-.03	.11	.17	.03	-.03
Subiculum Body	-.13	-.09	.15	-.02	.17	.15	-.03	-.05
CA1 Head	-.22	-.04	.12	-.02	.25	.23	.13	-.07
CA1 Body	-.13	-.07	-.04	-.04	-.03	.08	-.09	.07
CA3 Head	-.20	.01	.06	.01	.20	.26	.13	-.03
CA3 Body	-.06	.00	-.05	-.15	.11	.10	-.05	-.17
CA4 Head	-.07	-.07	.03	.03	.20	.26	.02	.02
CA4 Body	-.05	.01	-.06	-.10	.04	.07	-.08	-.20
GC ML DG Head	-.08	-.07	.03	.00	.12	.22	.03	.02
GC ML DG Body	-.20	-.06	.11	.00	.16	.04	.09	-.23
Molecular Layer HP Head	-.22	-.09	.07	-.01	.24	.26	.11	-.06
Molecular Layer HP Body	-.19	-.06	.11	-.05	.19	.09	.03	-.12

Fimbria	-.06	-.09	-.06	.03	.30	.19	.12	-.11
Hippocampal Fissure	-.24	-.31	.11	.23	.07	.11	-.03	.13
Hippocampal Tail	-.23	-.04	.23	-.01	.14	.24	.11	-.08
HATA	-.26	-.12	.18	.11	.04	.21	.19	.03

Table 3: Correlation coefficient for bivariate correlations between Hippocampal Subfield Dice Coefficients and subject-level covariates: interscan interval (ISI), MRI Quality (Difference Score; MRIQ), Sex, and Age. These were completed for left hemisphere (LH) and right hemisphere (RH). Correlations with $p<.05$ are highlighted in yellow. Subfield Abbreviations include: Cornu Ammonis (CA), Granule Cell and Molecular Layer of Dentate Gyrus (GC-ML-DG); Hippocampus-Amygdala-Transition-Area (HATA); Hippocampal Parcellation (HP).

Table 4.

Region	ISI r DICE		MRIQ r DICE		Sex r DICE		Age r DICE	
	LH	RH	LH	RH	LH	RH	LH	RH
Lateral Nucleus	-.14	-.03	.04	.02	.23	.17	.03	-.04
Basal Nucleus	-.13	.06	.00	-.06	.17	.19	.09	-.06
Central Nucleus	-.11	-.18	-.05	.13	.14	.04	.03	-.08
Medial Nucleus	-.19	-.30	.03	.19	.13	.09	.09	.16
Cortical Nucleus	-.26	.04	.12	.00	.19	.11	.25	-.00
Accessory Basal Nucleus	-.14	.11	.16	-.05	.18	.21	.14	-.02
Paralaminar Nucleus	.06	-.10	-.17	-.02	.11	.05	-.03	-.09
Corticoamygdaloid Transition	-.22	.00	.19	-.00	.09	.25	.20	.01
Anterior Amygdaloid Area	-.16	-.01	-.10	-.01	.06	.10	.18	.10

Table 4: Correlation coefficient for bivariate correlations between Hippocampal Subfield Dice Coefficients and subject-level covariates: interscan interval (ISI), MRI Quality (Difference Score; MRIQ), Sex, and Age. These were completed for left hemisphere (LH) and right hemisphere (RH). Correlations with $p<.05$ are highlighted in yellow.

Figure 1.

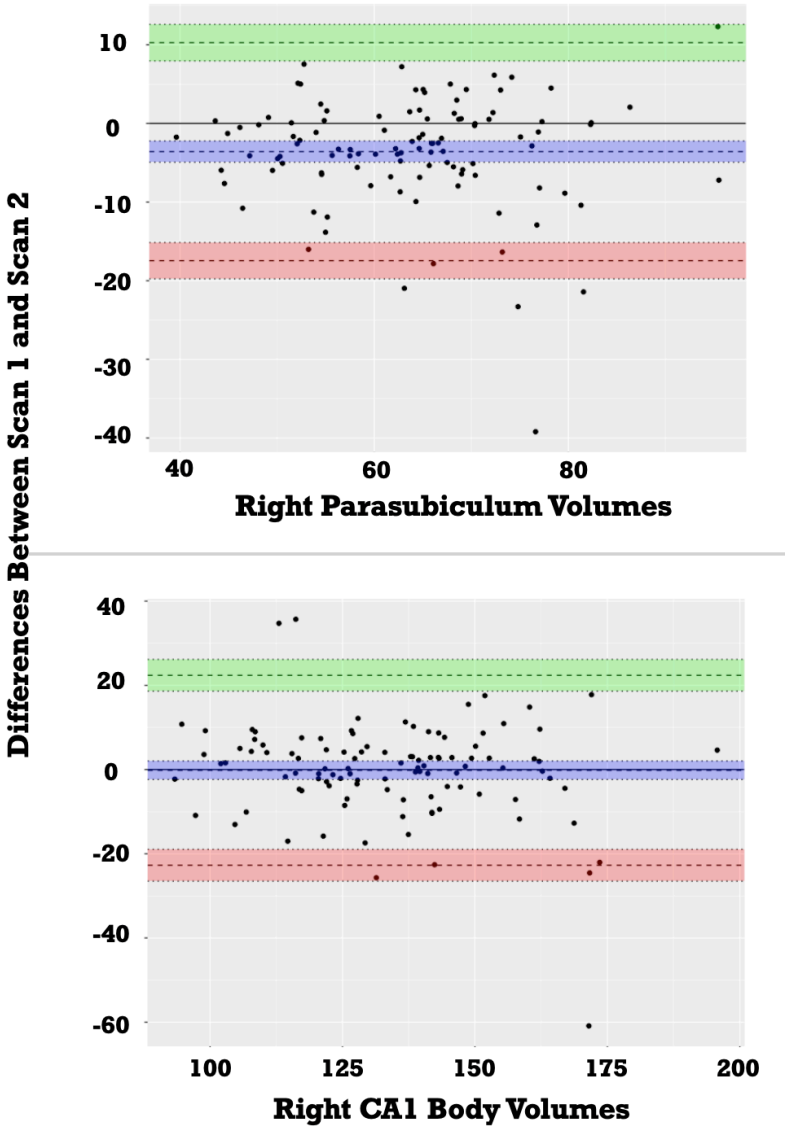

Figure 1: Bland–Altman plots of volume difference estimation for Scan 1 and Scan 2 for the right Parasubiculum (top) and right CA1 Body (bottom). The horizontal axis shows raw volumes for each participant for each region (Scan 1), while the vertical axis shows the difference (Scan 2 – Scan 1). The blue band is mean difference between scans, while the green band is +1SD and the red band is -1SD difference. Of note, the right Parasubiculum has a fair degree of bias (as a proportion of the region's volumes; 5.639%), while right CA1 Body has a relatively lower degree of bias (-0.130%).

Figure 2.

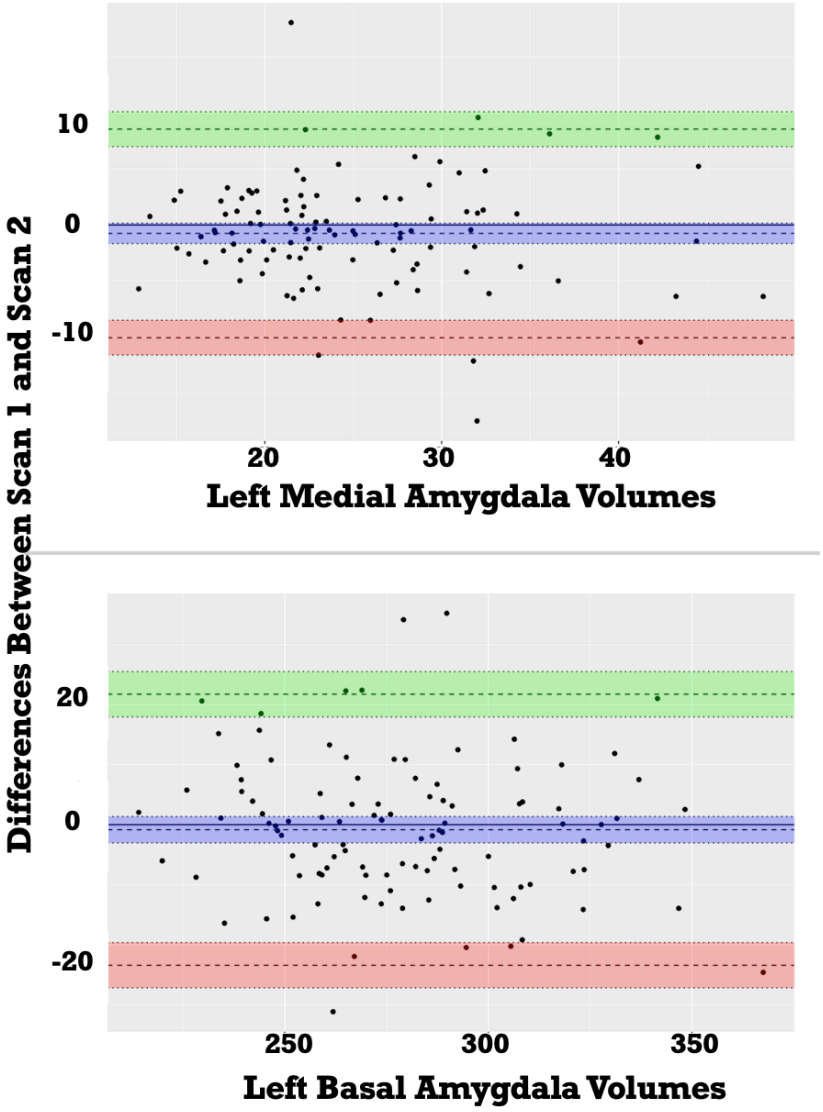

Figure 2: Bland–Altman plots of volume difference estimation for Scan 1 and Scan 2 for the Left Medial Amygdala (top) and Left Basal Amygdala (bottom). The horizontal axis shows raw volumes for each participant for each region (Scan 1), while the vertical axis shows the difference (Scan 2 – Scan 1). The blue band is mean difference between scans, while the green band is +1SD and the red band is -1SD difference. Of note, the Left Medial Amygdala has a fair degree of bias (as a proportion of the region's volumes; -3.004%), while Left Basal Amygdala has a relatively lower degree of bias (-0.216%).

Figure 3.

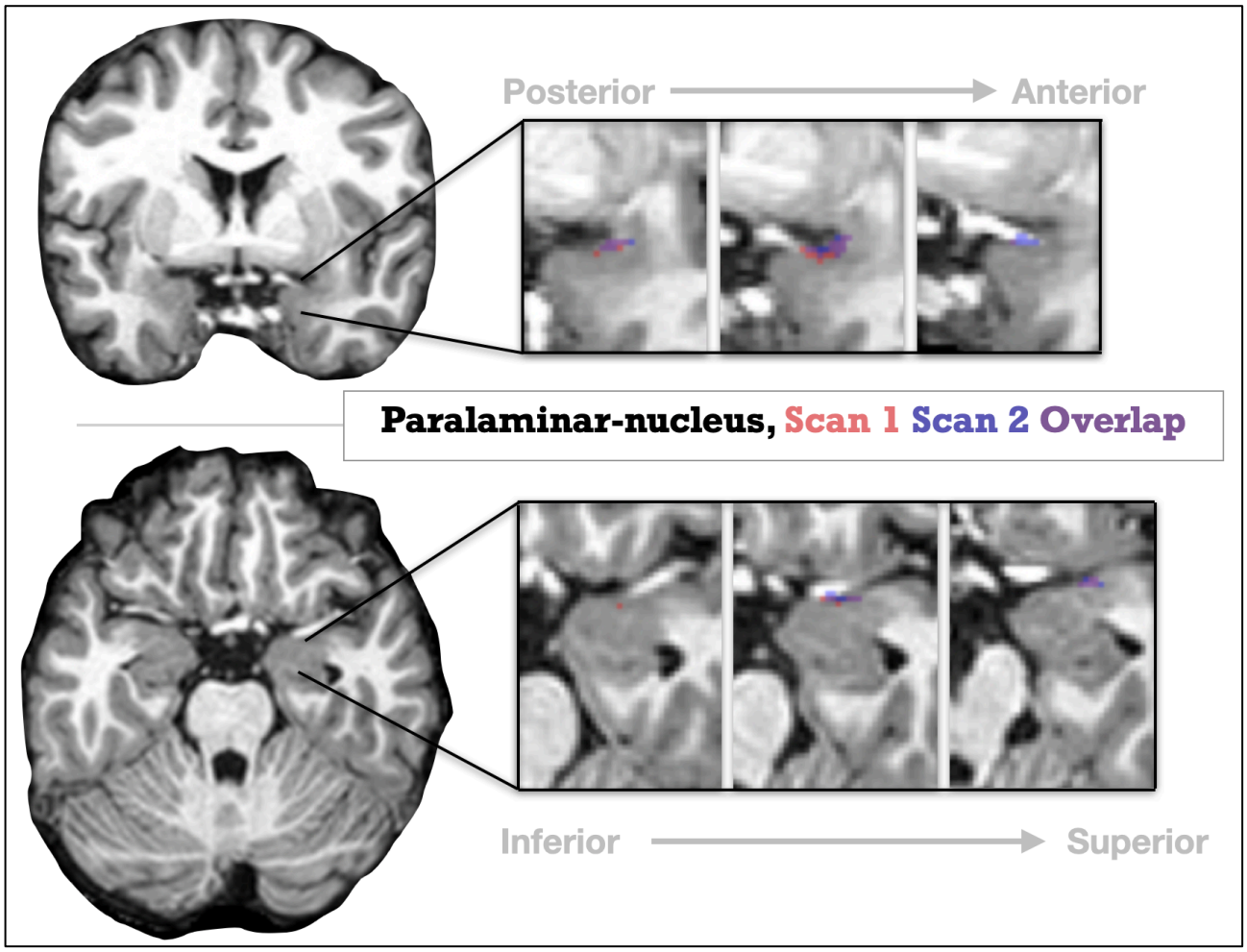

Figure 3: Graphic depictions showing coronal (top) and axial (bottom) views with magnified depictions of the segments of the Paralaminer nucleus from the repeated scans (Scan 1 shown in Red; Scan 2 shown in Purple). The anatomical (T1w) image underlaid is the unbiased subject template from an example 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 4.

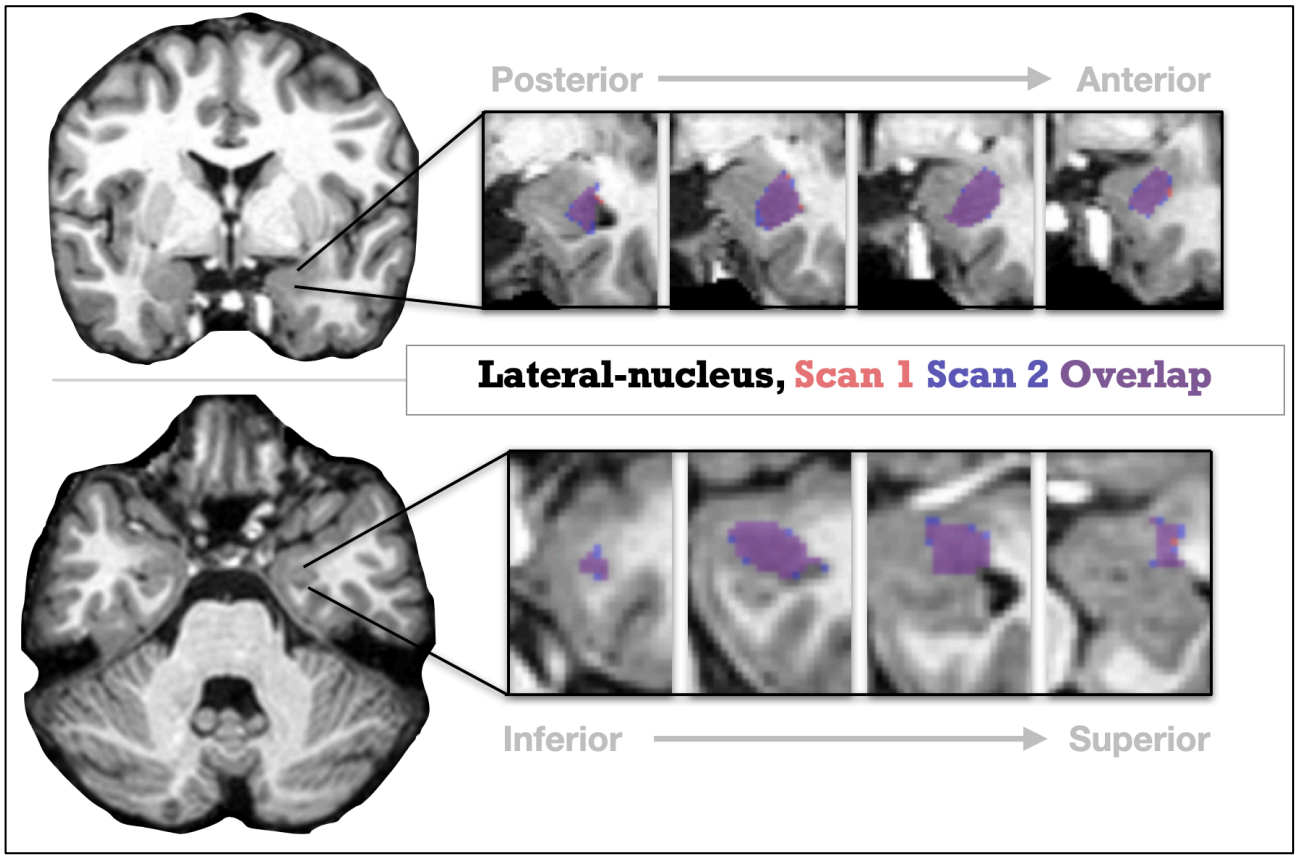

Figure 4: Graphic depictions showing coronal (top) and axial (bottom) views with magnified depictions of the segments of the Lateral nucleus from the repeated scans (Scan 1 shown in Red; Scan 2 shown in Purple). The anatomical (T1w) image underlaid is the unbiased subject template from an example participant. 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.