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

Isabella Kahhale[1], 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

Murdoch Building

3420 Forbes Ave, 6th Floor

Pittsburgh, PA 15260

Telephone: 412-383-3250

Email: jamie.hanson@pitt.edu

Abstract:

On-going, large-scale neuroimaging initiatives can aid in uncovering neurobiological causes and correlates of poor mental health, disease pathology, and many other important conditions. As projects grow in scale with hundreds, even thousands, of individual participants and scans collected, quantification of brain structures by automated algorithms is becoming the only *truly* tractable approach. Here, we assessed the spatial and numerical reliability for newly deployed automated segmentation of hippocampal subfields and amygdala nuclei in FreeSurfer 7. In a sample of participants with repeated structural imaging scans (*N=923*), we found numerical reliability (as assessed by intraclass correlations, ICCs) was reasonable: ~95% of hippocampal subfields have "excellent" numerical reliability (ICCs$\geq$0.90), however, only 67% of amygdala subnuclei met this same threshold. Spatial reliability was similarly reasonable, with 58% of hippocampal subfields and 44% of amygdala subnuclei having Dice coefficients$\geq$0.70. Notably, multiple regions had poor numerical and/or spatial reliability. We also examined correlations between spatial reliability and person-level factors (e.g., participant age; T1 image quality). Both sex and image scan quality 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; FreeSurfer 7.1*

1. Introduction

A large body of research has persistently emphasized 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 (Campbell et al., 2004; Hamilton et al., 2008; Stanfield et al., 2008; Yang et al., 2012). 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; Hanson et al., 2015; MacQueen et al., 2003; Malykhin et al., 2008; von Gunten & Ron, 2004; Yucel et al., 2007). These approaches were at the time necessary to obtain reliable and valid measures of the size of these key brain areas, but hand-tracing is often exceedingly time intensive. As work in this space has continued, 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. Researchers are now able to leverage ever-improving computational algorithms to automatically segment structural images into their component anatomical structures (Madan & Kensinger, 2017). These approaches represent a scalable and less demanding method to 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 the 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 reflects 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). Results considering validity have been more mixed. 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 reasonable Dice coefficients for the hippocampus, but lower performance on the amygdala (Hippocampus Dice coefficient=0.82; Amygdala 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 the following regions: 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). Most of the past structural neuroimaging work has combined all these regions, using measures of whole hippocampal volume. This may mean a loss of specificity regarding associations with 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 considering 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 combined ultra-high-resolution T1-weighted scans of post-mortem samples with subfields of the hippocampus segmented by hand, to develop an automated algorithm. With this algorithm, there appears to be good numerical reliability and slightly lower spatial reliability for these segments, mirroring the reliability work focusing on the whole hippocampus. Numerical reliability and ICCs are focused on the consistent overall volume size (as indexed by the number of voxels in a region), whereas spatial reliability and the calculation of Dice coefficients assess 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 in fact spatially divergent (and may not be the same brain area). 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, similarly, has its own subdivisions and the reliability of these subdivisions are still unclear. The FreeSurfer team also 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 (Di Martino et al., 2014) and those at risk for Alzheimer's disease (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, direct assessment of numerical and spatial reliability for amygdala subdivisions is limited. Quattrini and colleagues (2020) examined these segments in a modest cohort of individuals (total *N=133*) and found reasonable reliability for larger subdivisions (>200 mm^3 for the amygdala; >300 mm^3 for the hippocampus). This work, however, aggregated across 17 research sites and multiple MRI vendors, deployed a dated version of the software (Freesurfer 6.0), and often acquired repeated imaging scans across weeks and months. Given these limitations, the consistency of these segments is an open question, and it is still unclear whether this fine-grained separation is consistent in the areas that the algorithm is automatically dividing and outputting. Such gaps are critical to fill 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, we seek to provide an in-depth examination of reliability, both numerically and spatially, for FreeSurfer derived hippocampal and amygdala subdivisions. We leverage a public-access dataset of repeated structural scans with a large sample size (*N=928*). In addition to this first-order goal, we also wanted to consider whether person-level (e.g., age, sex) and MR-acquisition (e.g., 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., 2021; 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, e.g., tests of amygdala subdivisions in depressed vs. non-depressed groups).

2. Methods

Participants

Data from an open-access neuroimaging initiative, the Amsterdam Open MRI Collection (AOMIC) (Snoek et al., 2021) was used to investigate numerical and spatial reliability of FreeSurfer's amygdala and hippocampal subregion segmentation algorithms. AOMIC includes structural and functional neuroimaging scans from participants, repeating scans in the same session to see the stability of MRI-based metrics. For this work, data from 928 participants (Average Age=22.08, Standard Deviation =1.88) was examined. The majority of participants (n = 913, 98% of the sample) had three T1-weighted MR images collected in the same scanning session, while a small subgroup of participants (n = 15, ~2% of the sample) had two T1-weighted scans. Additionally, 5 subjects were removed for extreme outlier values suggesting a processing error. The final analytic was 923 participants. All repeated MRI scans were acquired with the same imaging parameters (noted below).

MRI Scan Parameters

MR images were acquired with a Phillips 3T scanner ("*Intera*" version) at the University of Amsterdam. T1-weighted MR images were acquired using a sagittal 3D-MPRAGE sequence (TR/TE = 8.1ms/3.7ms, 1mm^3 voxel, matrix size = 64x64). Additional details

about the scanning parameters are described by Snoek and colleagues (2021). 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 (Avesani et al., 2019; Pestilli, 2018). FreeSurfer is a widely documented and freely available morphometric processing tool suite (http://surfer.nmr.mgh.harvard.edu). The technical details of this software suite 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 hippocampal 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 temporal consistency and reduce potential source of bias (e.g., misregistration) (Reuter et al., 2010, 2012). Cross-sectional pipelines were applied to the three scans 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 output for each subdivision. Each participant's T1-weighted scan was then transformed to a common space using FMRIB's Linear Image Registration Tool (degrees of freedom = 6; registering the 2nd and 3rd scans to the participant's 1st scan). This transformation matrix was then saved and applied to each volume's labelled output for hippocampal and amygdala subdivisions using a nearest neighbour interpolation; these transformed hippocampal and amygdala subdivisions were then used in spatial reliability analysis.

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 one quality metric for each scan, we

generated three values that represent the difference between two scans (i.e., Scan 1 – Scan 2; Scan 1 – Scan 3; Scan 2 – Scan 3). After taking the absolute value of each of these difference scores, we then averaged them together and used this as a measure of aggregate image quality.

Derivation of Reliability Measures

To assess the reliability of numerical volumes 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 using the function blandr.statistics from the package '*blandr'* (Datta, 2017). In this approach, the mean differences ("*bias*") between the FreeSurfer outputs

(comparing the first and second scan, the first and third scan, and the second and third scan) were first calculated and presented as a portion of the mean volume. We took the absolute value of each of these three values and averaged them together to represent the average Bland-Altman metric across the three scans for a given brain region. Bland-Altman plots were also constructed for a small number of 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. Dice coefficients were averaged across the three scans for each brain region to obtain an overall metric of spatial reliability (e.g., one Dice value for the Left Lateral Nucleus, one Dice value for the Right Lateral Nucleus). 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 overall reliability metrics were calculated, we examined person-level (e.g., age, sex) and MR-acquisition (e.g., 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). Finally, image quality has been shown to have a significant effect on brain volume measurements (Gilmore et al., 2021). 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, 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 average difference in quality between the three scans (described in *Automated MRI Image Quality Assessment*) 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 all of the subfields were at least in the “good” range (ICC=0.75-0.90). See Table 1 for values from the 19 subfield segmentations in each hemisphere. Bland-Altman bias indicated some variability with differences between scans, as a portion of that structure’s volume, ranging from 0.078-1.198%. See Figure 1 for a density plot of the average difference in volume estimation across three scans for two hippocampal subfields.

---- Table 1 Approximately Here ---

---- Figure 1 Approximately Here ---

Using Dice coefficients as metrics of spatial reliability, results became a bit more variable with 11 areas showing “excellent” spatial reliability, 7 areas showing “good” spatial reliability, and one area (left Hippocampal fissure) showing poor spatial reliability (Dice Coefficient<0.5.). See Figure 2 for a plot of all Hippocampal Dice Coefficient values and Figure 3 for an example of regions with acceptable spatial reliability (parasubiculum) and poor spatial reliability (hippocampal fissure).

---- Figure 2 Approximately Here ---

---- Figure 3 Approximately Here ---

Amygdala Reliability

Within the amygdala, the numerical reliability was “excellent” for about 67% of the regions (ICC > .90), while the remainder of the regions were in the “good” range (ICC = 0.75-0.90) (*see Table 2*). Bland-Altman bias values were somewhat variable with a range of 0.058-1.563%. See Figure 4 for a density plot of the average difference in volume estimation across three scans for two amygdala subnuclei.

---- Table 2 Approximately Here ---

---- Figure 4 Approximately Here ---

Regarding spatial reliability, several areas demonstrated good reliability (≥.7) including the lateral, basal, and accessory basal subnuclei (See *Table 2*). There were, however, areas

with poor spatial reliability (Dice Coefficients = 0.30-0.4, See *Figure 5* for a plot of all Amygdala Dice Coefficient values). These problems were primarily seen in the Medial and Paralaminar Nuclei. Figure 6 displays a depiction of the Lateral nucleus, an area with acceptable spatial reliability, and the Paralaminar nucleus, an area with poor spatial reliability.

---- *Figure 5 Approximately Here* ---

---- *Figure 6 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 Table 3. Differences in image quality and participant sex were significantly and negatively related to volumes in a majority of the hippocampal subfields at the $p < 0.01$ level (shown in Table 3). Age was not significantly correlated with any hippocampal subfield volumes.

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

Correlations between spatial reliability and subject-level variables for the amygdala nuclei are reported in Table 4. Image quality was significantly and negatively related to a small number of regions including the lateral nucleus, the right basal nucleus, the corticoamygdaloid transition, and the right anterior amygdaloid area (at $p<0.01$).

The spatial reliability of a minority of regions was also significantly and associated with sex and age.

---- Table 4 Approximately Here ---

4. Discussion

In this paper, we assessed the numerical and spatial reliability of FreeSurfer's hippocampal and amygdala subdivision segmentation algorithms. The ICC's, serving as our indicator of numerical reliability, were reasonable (hippocampal subfields: 0.887-0.979; amygdala nuclei: 0.832-0.964), 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, 5 regions across the hippocampus and amygdala showed $\geq$1% bias in volume from one scan to the next. This is concerning given that individuals with dementia (i.e., Alzheimer's disease) or recurrent mental health issues (i.e., depression) often only differ 1-5% from control groups in subcortical volumes (e.g., Jack et al., 2005; Logue et al., 2018; Schmaal et al., 2016). The Dice coefficients, serving as our indicator of spatial reliability, were reasonable, though 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 ($\leq$.5) 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 potential drivers of the substandard performance in this domain. First, these areas are fairly small 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 $\geq$ 0.765 Dice Coefficients; Parasubiculum $\geq$ 0.712 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, several 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, 1999; Watson et al., 1992, as cited in Entis et al., 2012). Many past reports using manual tracing 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 collaboration 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 the segmentation of hippocampal subregions (eg., Olsen et al., 2019; Wisse et al., 2017; Yushkevich et al., 2015). This and other related work have suggested important ways to validate automatic segmentation, including not only comparison to manual delineations, but also replicating known disease effects (e.g., Mueller et al., 2018). 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 & 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 et al., 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). There is, however, ongoing debate about specific best practices, as even established guidelines for MRI acquisition or landmark in in-vivo data may present additional unforeseen challenges (e.g., Special hippocampal acquisitions providing incomplete coverage of target structure; In-vivo MRI does not supply enough features to define many hippocampal subfield boundaries).

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 automated amygdala nuclei segmentation to hand-tracing. Reliable methods exist for expert manual segmentation of the amygdala (Aghamohammadi-Sereshki et al., 2018; Entis et al., 2012); 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.

In considering our results, it is important to note a few potential limitations of our work. First, 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. Second, we only used T1-weighted scans in FreeSurfer, but additional MRI

volumes (e.g., T2-weighted) from the same subjects would likely yield a more reliable segmentation. FreeSurfer’s developers have worked to allow the amygdala and hippocampal subdivision routines to accept high-resolution T2-weighted volumes, and this should be investigated in future work. Third, the sample is a rather homogenous group of individuals and may not represent the greater population. All participants were recruited from the University of Amsterdam, with limited race and ethnic variability. Similarly, all participants were neurotypical young adults in a constrained age range (Mean=22.08+/-1.88). 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.

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.

Data Availability Statement: Neuroimaging data used in our analyses were sourced from the Amsterdam Open MRI Collection (AOMIC, Snoek et al., 2021). 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 brainlife.io.

Conflicts of interest/Competing interests: The authors have no conflicts of interest to disclose.

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. https://doi.org/10.1007/s002340050644

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. https://doi.org/10.1016/j.neuroimage.2017.03.016

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*(1), 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

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*, 116563. https://doi.org/10.1016/j.neuroimage.2020.116563

Caldwell, J. Z. K., Armstrong, J. M., Hanson, J. L., Sutterer, M. J., Stodola, D. E., Koenigs, M., Kalin, N. H., Essex, M. J., & Davidson, R. J. (2015). Preschool Externalizing

Behavior Predicts Gender-Specific Variation in Adolescent Neural Structure. *PLOS ONE*, *10*(2), e0117453. https://doi.org/10.1371/journal.pone.0117453

Campbell, S., Marriott, M., Nahmias, C., & MacQueen, G. M. (2004). Lower Hippocampal Volume in Patients Suffering From Depression: A Meta-Analysis. *American Journal of Psychiatry*, *161*(4), 598–607. https://doi.org/10.1176/appi.ajp.161.4.598

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.

Convit, A. (1999). MRI volume of the amygdala: A reliable method allowing separation from the hippocampal formation. *Psychiatry Research: Neuroimaging*, *90*(2), 113–123. https://doi.org/10.1016/S0925-4927(99)00007-4

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*. https://doi.org/10.13140/RG.2.2.16267.44321

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *9*(2), 179–194.

Datta, D. (2017). blandr: A Bland-Altman Method Comparison Package for R. Zenodo. *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, *52*(3), 11. https://doi.org/10.5281/zenodo. 824514.

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

Di Martino, A., 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. (2014). 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

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. *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*, S69–S84. https://doi.org/10.1016/j.neuroimage.2004.07.016

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *9*(2), 195–207.

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.

Gamer, M., & Lemon, J. (2012). *Package "irr."* 32.

Gaser, C., & Kurth, F. (2017). Manual computational anatomy toolbox-CAT12. *Structural Brain Mapping Group at the Departments of Psychiatry and Neurology, University of Jena*, 69.

Gilmore, A. D., Buser, N. J., & Hanson, J. L. (2021). Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. *Brain Informatics*, *8*(1), 1–15. https://doi.org/10.1186/s40708-021-00128-2

Gunten, A., Fox, N., 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 Neuropsychiatry*, *12*(4), 6. https://doi.org/10.1001/jama.1992.03490110111047

Guzman, S. J., Schlogl, 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

Hamilton, J. P., Siemer, M., & Gotlib, I. H. (2008). Amygdala volume in major depressive disorder: A meta-analysis of magnetic resonance imaging studies. *Molecular Psychiatry*, *13*(11), 993–1000. https://doi.org/10.1038/mp.2008.57

Hanson, J. L., Nacewicz, B. M., Sutterer, M. J., Cayo, A. A., Schaefer, S. M., Rudolph, K. D., Shirtcliff, E. A., Pollak, S. D., & Davidson, R. J. (2015). Behavioral Problems After Early Life Stress: Contributions of the Hippocampus and Amygdala. *Biological Psychiatry*, *77*(4), 314–323. https://doi.org/10.1016/j.biopsych.2014.04.020

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

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

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. (2016). Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. *NeuroImage*, *141*, 542–555. https://doi.org/10.1016/j.neuroimage.2016.07.020

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., 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., … ADNI Study. (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

Jack, C. R., Shiung, M. M., Weigand, S. D., O'Brien, P. C., Gunter, J. L., Boeve, B. F., Knopman, D. S., Smith, G. E., Ivnik, R. J., Tangalos, E. G., & Petersen, R. C. (2005). Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. *Neurology*, *65*(8), 1227–1231. https://doi.org/10.1212/01.wnl.0000180958.22678.91

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., 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*, 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

Logue, M. W., van Rooij, S. J. H., Dennis, E. L., Davis, S. L., Hayes, J. P., Stevens, J. S., Densmore, M., Haswell, C. C., Ipser, J., Koch, S. B. J., Korgaonkar, M., Lebois, L. A. M., Peverill, M., Baker, J. T., Boedhoe, P. S. W., Frijling, J. L., Gruber, S. A., Harpaz-Rotem, I., Jahanshad, N., … Morey, R. A. (2018). Smaller Hippocampal Volume in Posttraumatic Stress Disorder: A Multisite ENIGMA-PGC Study: Subcortical Volumetry Results From Posttraumatic Stress Disorder Consortia.

Biological Psychiatry, *83*(3), 244–253. https://doi.org/10.1016/j.biopsych.2017.09.006

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. *PNAS*, *100*(3), 1387–1392.

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

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.

Merboldt, K.-D., Fransson, P., Bruhn, H., & Frahm, J. (2001). Functional MRI of the Human Amygdala? *NeuroImage*, *14*(2), 253–257. https://doi.org/10.1006/nimg.2001.0802

Morey, R. A., Clarke, E. K., Haswell, C. C., Phillips, R. D., Clausen, A. N., Mufford, M. S., Saygin, Z., Wagner, H. R., LaBar, K. S., 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., … Yoash-Gantz, R. E. (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*, NA-NA. https://doi.org/10.1002/hbm.20973

Mueller, S. G., Yushkevich, P. A., Das, S., Wang, L., Van Leemput, K., Iglesias, J. E., Alpert, K., Mezher, A., Ng, P., Paz, K., & Weiner, M. W. (2018). Systematic comparison of different techniques to measure hippocampal subfield volumes in ADNI2. *NeuroImage: Clinical*, *17*, 1006–1018. https://doi.org/10.1016/j.nicl.2017.12.036

Nacewicz, B. M., Dalton, K. M., Johnstone, T., Long, M. T., McAuliff, E. M., Oakes, T. R., Alexander, A. L., & 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. https://doi.org/10.1001/archpsyc.63.12.1417

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. C., Amunts, K., Augustinack, J. C., Bakker, A., Bender, A. R., Berron, D., Boccardi, M., Bocchetta, M., Burggren, A. C., Chakravarty, M. M., Chételat, G., Flores, R., DeKraker, J., Ding, S., Geerlings, M. I., … Hippocampal Subfields Group. (2019). Progress update from the hippocampal subfields group. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *11*(1), 439–449. https://doi.org/10.1016/j.dadm.2019.04.001

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. https://doi.org/10.1177/1077559519839491

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 Biology*, *16*(4), 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

Quattrini, G., Pievani, M., Jovicich, J., Aiello, M., Bargalló, N., Barkhof, F., Bartres-Faz, D., Beltramello, A., Pizzini, F. B., Blin, O., Bordet, R., Caulo, M., Constantinides, M.,

Didic, M., Drevelegas, A., Ferretti, A., Fiedler, U., Floridi, P., Gros-Dagnac, H., … Marizzoni, M. (2020). Amygdalar nuclei and hippocampal subfields on MRI: Test-retest reliability of automated volumetry across different MRI sites and vendors. *NeuroImage*, *218*, 116932. https://doi.org/10.1016/j.neuroimage.2020.116932

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. https://doi.org/10.1016/j.neuroimage.2003.12.048

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

Schmaal, L., Veltman, D., Erp, T., Samann, P., & Frodl, T. (2016). Subcortical brain alterations in major depressive disorder: Findings from the ENIGMA Major Depressive Disorder working group. *Molecular Psychiatry*, *21*(6), 806–812.

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

Snoek, L., van der Miesen, M. M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., & Steven Scholte, H. (2021). The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data*, *8*(1), 85. https://doi.org/10.1038/s41597-021-00870-6

Stanfield, A. C., McIntosh, A. M., Spencer, M. D., Philip, R., Gaur, S., & Lawrie, S. M. (2008). Towards a neuroanatomy of autism: A systematic review and meta-analysis of structural magnetic resonance imaging studies. *European Psychiatry*, *23*(4), 289–299. https://doi.org/10.1016/j.eurpsy.2007.05.006

Swanson, L. W., & Petrovich, G. D. (1998). What is the amygdala? *Trends in Neurosciences*, *21*(8), 323–331. https://doi.org/10.1016/S0166-2236(98)01265-X

von Gunten, A., & 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

Whelan, C. D., Hibar, D. P., van Velzen, L. S., Zannas, A. S., Carrillo-Roa, T., McMahon, K., Prasad, G., Kelly, S., Faskowitz, J., deZubiracay, 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.

https://doi.org/10.1016/S0730-725X(02)00563-5

Wisse, L. E. M., Daugherty, A. M., Olsen, R. K., Berron, D., Carr, V. A., Stark, C. E. L., Amaral, R. S. C., Amunts, K., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., Bocchetta, M., Burggren, A., Chakravarty, M. M., Chupin, M., Ekstrom, A., de Flores, R., Insausti, R., … for the Hippocampal Subfields Group. (2017). A harmonized segmentation protocol for hippocampal and parahippocampal subregions: Why do we need one and what are the key goals?: A HARMONIZED HIPPOCAMPAL SUBFIELD PROTOCOL: KEY GOALS AND IMPACT. *Hippocampus*, *27*(1), 3–11. https://doi.org/10.1002/hipo.22671

Wonderlick, J., Ziegler, D., Hosseinivarnamkhasti, P., Locascio, J., Bakkour, A., Vanderkouwe, A., Triantafyllou, C., Corkin, S., & Dickerson, B. (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., Williams, S. C. R., & O'daly, O. (2018). Test–retest reliability and longitudinal analysis of automated hippocampal subregion volumes in healthy ageing and A lzheimer's disease populations. *Human Brain Mapping*, *39*(4), 1743–1754. https://doi.org/10.1002/hbm.23948

Yang, J., Pan, P., Song, W., Huang, R., Li, J., Chen, K., Gong, Q., Zhong, J., Shi, H., & Shang, H. (2012). Voxelwise meta-analysis of gray matter anomalies in Alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation. *Journal of the Neurological Sciences*, *316*(1–2), 21–29. https://doi.org/10.1016/j.jns.2012.02.010

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. C., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., Bocchetta, M., Burggren, A. C., Carr, V. A., Chakravarty, M. M., Chételat, G., Daugherty, A. M., Davachi, L., Ding, S.-L., Ekstrom, A., Geerlings, M. I., Hassan, A., Huang, Y., Iglesias, J. E., … Zeineh, M. M. (2015). Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. *NeuroImage*, *111*, 526–541. https://doi.org/10.1016/j.neuroimage.2015.01.004

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. https://doi.org/10.1109/42.363096

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M. C., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index1. *Academic Radiology*, *11*(2), 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8

Table 1.

Region	ICC LH	ICC RH	Dice LH	Dice RH	Bias as POV LH	Bias Range LH	Bias as POV RH	Bias Range RH
Parasubiculum	0.929	0.946	0.713	0.710	0.356%	0.000 – 45.834%	0.410%	0.008 – 32.920%
Presubiculum Head	0.924	0.936	0.793	0.791	0.212%	0.001 – 29.992%	0.078%	0.002 – 23.345%
Presubiculum Body	0.960	0.963	0.800	0.791	1.057%	0.000 – 41.021%	0.626%	0.001 – 33.210%
Subiculum Head	0.961	0.959	0.777	0.773	0.130%	0.001 – 27.278%	0.085%	0.002 – 25.029%
Subiculum Body	0.961	0.964	0.823	0.826	0.100%	0.004 – 31.485%	0.289%	0.002 – 16.760%
CA1 Head	0.971	0.979	0.819	0.825	0.106%	0.000 – 22.307%	0.260%	0.000 – 16.274%
CA1 Body	0.948	0.970	0.752	0.780	0.127%	0.001 – 50.541%	0.465%	0.001 – 29.594%
CA3 Head	0.952	0.969	0.663	0.677	0.533%	0.000 – 24.448%	0.742%	0.000 – 22.072%
CA3 Body	0.933	0.950	0.598	0.627	0.413%	0.000 – 55.834%	0.973%	0.001 – 27.936%
CA4 Head	0.966	0.966	0.794	0.798	0.551%	0.000 – 20.589%	0.494%	0.000 – 21.287%
CA4 Body	0.931	0.938	0.768	0.780	0.541%	0.002 – 29.764%	0.625%	0.002 – 20.289%
GC ML DG Head	0.965	0.973	0.618	0.626	0.554%	0.001 – 20.329%	0.542%	0.002 – 18.378%
GC ML DG Body	0.940	0.942	0.593	0.600	0.500%	0.000 – 21.930%	0.503%	0.001 – 27.053%
Molecular Layer HP Head	0.971	0.976	0.690	0.692	0.206%	0.001 – 18.600%	0.240%	0.001 – 14.279%
Molecular Layer HP Body	0.954	0.961	0.632	0.647	0.495%	0.001 – 25.208%	0.558%	0.000 – 18.751%
Fimbria	0.936	0.942	0.681	0.679	0.639%	0.008 – 78.157%	1.198%	0.004 – 57.572%
Hippocampal Fissure	0.887	0.888	0.498	0.511	0.443%	0.001 – 41.512%	0.586%	0.000 – 45.252%
Hippocampal Tail	0.954	0.968	0.882	0.890	0.315%	0.000 – 50.169 %	0.393%	0.001 – 18.664%
HATA	0.936	0.945	0.760	0.770	1.049%	0.000 – 34.712%	0.216%	0.007 – 26.125%

Whole Hippocampal Body	0.959	0.966			0.365%	0.000 – 27.935%	0.435%	0.004 – 15.245%
Whole Hippocampal Head	0.975	0.980			0.227%	0.001 – 19.277%	0.273%	0.002 – 13.619%
Whole Hippocampus	0.976	0.983			0.251%	0.001 – 21.668%	0.341%	0.000 – 12.473%

Table 1: Intraclass Correlation Coefficients (ICC), Dice Coefficients, Bland–Altman bias as a portion of a volume's structure (Bias as POV), and Bland-Altman bias ranges 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], 0.75-0.89 [good], 0.00-0.74 [poor]; Dice Coefficients: 0.70-1.00 [excellent], 0.50-.69 [good], 0-0.49 [poor]). We have also highlighted regions with >1% 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 Range LH	Bias as POV RH	Bias Range RH
Lateral Nucleus	0.964	0.956	0.876	0.868	0.108%	0.003 – 24.233%	0.200%	0.001 – 16.662%
Basal Nucleus	0.959	0.956	0.858	0.847	0.387%	0.001 – 24.086%	0.213%	0.001 – 20.245%
Central Nucleus	0.895	0.867	0.527	0.508	0.333 %	0.007 – 53.459%	0.331%	0.002 – 42.310%
Medial Nucleus	0.845	0.832	0.339	0.308	1.563%	0.004 – 73.504%	1.047%	0.005 – 70.814%
Cortical Nucleus	0.889	0.905	0.511	0.521	0.376%	0.001 – 68.640%	0.703%	0.001 – 29.931%
Accessory Basal Nucleus	0.957	0.961	0.848	0.833	0.224%	0.000 – 21.655%	0.647%	0.001 – 16.607%
Paralaminar Nucleus	0.941	0.946	0.394	0.383	0.143%	0.000 – 27.786%	0.199%	0.003 – 17.924%
Corticoamygdaloid Transition	0.939	0.949	0.698	0.684	0.588%	0.002 – 37.478%	0.704%	0.005 – 17.411%
Anterior Amygdaloid Area	0.901	0.858	0.597	0.583	0.058%	0.006 – 39.745%	0.541%	0.003 – 41.760%
Whole Amygdala	0.974	0.967			0.149%	0.002 – 21.679%	0.328%	0.001 – 13.837%

Table 2: Intraclass Correlation Coefficients (ICCs), Dice Coefficients, Bland–Altman bias as a portion of a volume's structure (Bias as POV), and Bland-Altman bias ranges 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], 0.75-0.89 [good], 0.00-0.74 [poor]; Dice Coefficients: 0.70-1.00 [excellent], 0.50-.69 [good], 0-0.49 [poor]). We have also highlighted regions with >1% bias as a portion of a volume's structure in yellow.

Table 3.

Region	MRIQ r Dice		Sex r Dice		Age r Dice	
	LH	RH	LH	RH	LH	RH
Parasubiculum	-0.13	-0.08	-0.03	-0.09	0.00	-0.05
Presubiculum Head	-0.13	-0.16	-0.06	-0.10	-0.04	-0.04
Presubiculum Body	-0.15	-0.16	-0.08	-0.08	-0.02	0.01
Subiculum Head	-0.13	-0.17	-0.06	-0.09	-0.03	-0.05
Subiculum Body	-0.09	-0.11	-0.06	-0.09	0.02	-0.03
CA1 Head	-0.13	-0.14	-0.07	-0.12	-0.06	-0.05
CA1 Body	-0.04	-0.07	-0.08	-0.13	0.01	-0.06
CA3 Head	-0.09	-0.07	-0.09	-0.12	0.01	-0.06
CA3 Body	-0.05	-0.08	-0.09	-0.10	0.01	-0.01
CA4 Head	-0.09	-0.13	-0.08	-0.09	-0.04	-0.07
CA4 Body	-0.04	-0.11	-0.09	-0.08	0.01	-0.04
GC ML DG Head	-0.10	-0.11	-0.07	-0.10	-0.03	-0.08
GC ML DG Body	-0.06	-0.10	-0.09	-0.10	0.01	-0.05
Molecular Layer HP Head	-0.13	-0.14	-0.06	-0.11	-0.05	-0.07
Molecular Layer HP Body	-0.07	-0.10	-0.08	-0.10	0.02	-0.05
Fimbria	-0.12	-0.17	-0.01	-0.05	0.02	-0.02
Hippocampal Fissure	-0.10	-0.17	-0.08	-0.13	-0.04	-0.05
Hippocampal Tail	-0.08	-0.14	-0.12	-0.11	-0.03	-0.01
HATA	-0.08	-0.10	-0.09	-0.12	0.00	-0.04

Table 3: Correlation coefficient for bivariate correlations between Hippocampal Subfield Dice Coefficients and subject-level covariates: MRI Quality (Difference Score; MRIQ), Sex, and Age. These were completed for left hemisphere (LH) and right hemisphere (RH). Correlations with $p<.01$ 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	MRIQ r Dice		Sex r Dice		Age r Dice	
	LH	RH	LH	RH	LH	RH
Lateral Nucleus	-0.12	-0.18	-0.10	-0.11	0.01	-0.03
Basal Nucleus	-0.07	-0.10	-0.05	-0.06	-0.03	-0.07
Central Nucleus	-0.05	0.01	0.04	0.01	-0.03	-0.10
Medial Nucleus	-0.07	-0.01	-0.09	-0.06	0.00	-0.01
Cortical Nucleus	0.00	-0.04	-0.06	-0.04	0.04	-0.02
Accessory Basal Nucleus	-0.05	-0.07	-0.06	-0.03	-0.03	-0.08
Paralaminar Nucleus	-0.03	-0.07	-0.01	-0.04	-0.02	-0.02
Corticoamygdaloid Transition	-0.12	-0.10	-0.08	-0.07	-0.02	-0.06
Anterior Amygdaloid Area	-0.07	-0.13	-0.09	-0.06	-0.02	-0.05

Table 4: Correlation coefficient for bivariate correlations between Hippocampal Subfield Dice Coefficients and subject-level covariates: MRI Quality (Difference Score; MRIQ), Sex, and Age. These were completed for left hemisphere (LH) and right hemisphere (RH). Correlations with $p<.01$ are highlighted in yellow.

Figure 1.

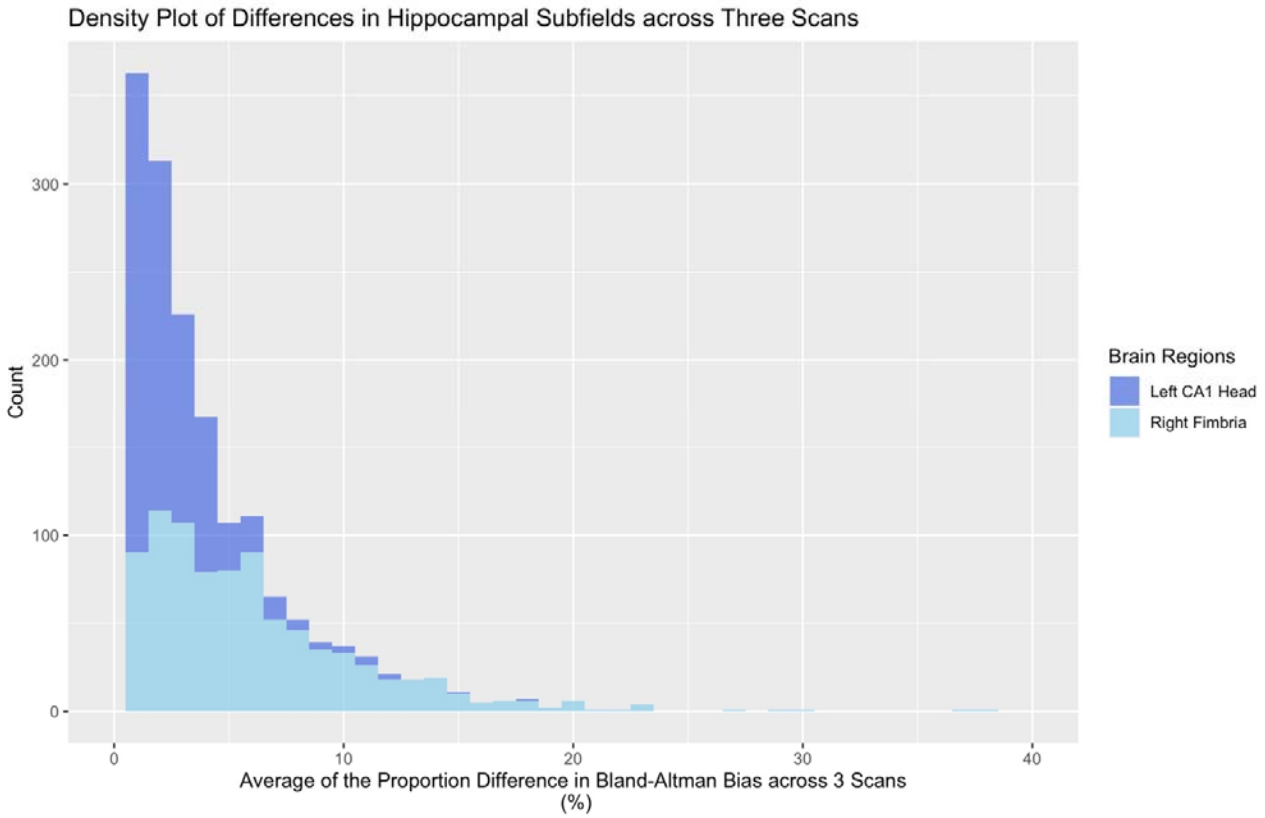

Figure 1: Bland–Altman plots of the average difference for volume estimation across subjects' three MRI scans for the Left Cornu Ammonis (CA) 1 Head (dark blue) and Right Fimbria (light blue). The horizontal axis indicates the average difference in Bland-Altman "*bias*" (difference between subregional volume output for different scans, as a proportion of a region's volume), while the vertical axis indicates the number of scans with a given value. Of note, the left CA1 Head has a low degree of mean bias (as a proportion of the region's volumes; 0.106%), while the right Fimbria has a fair degree of mean bias (1.198%).

Figure 2.

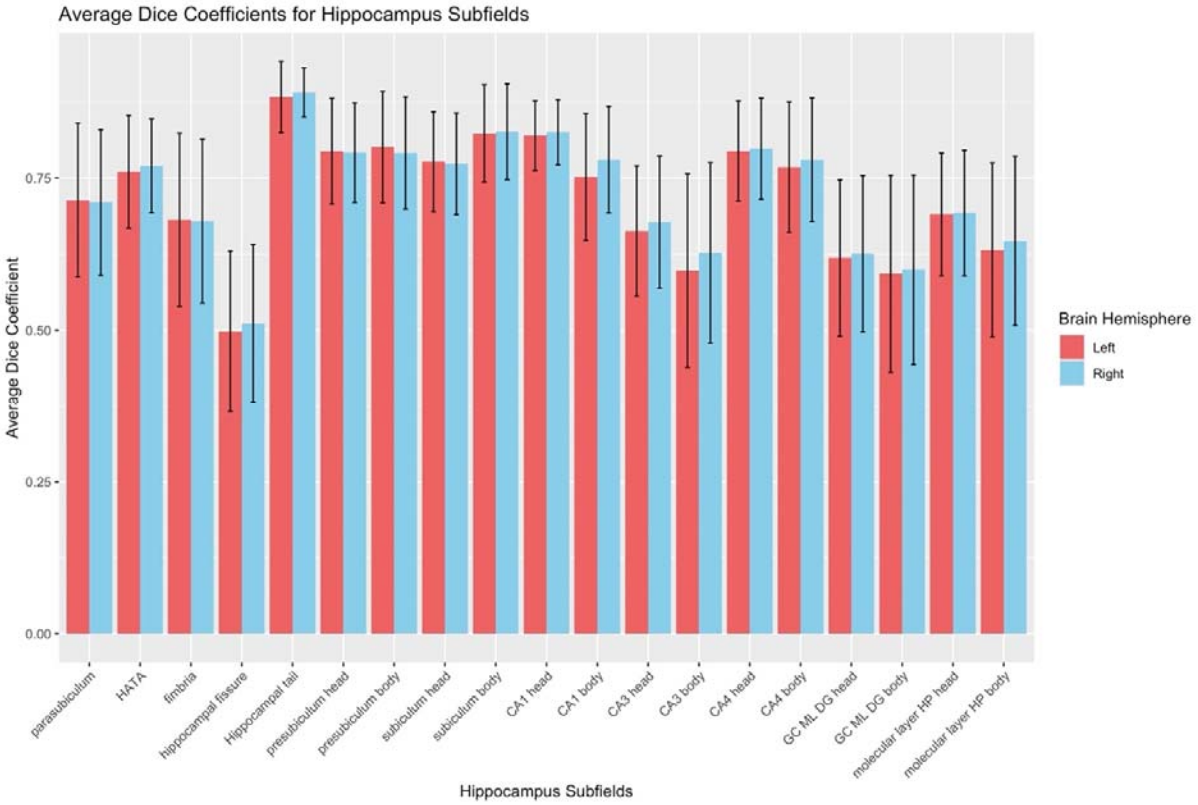

Figure 2. Hippocampal Dice Coefficient values for all hippocampal subfields. Error bars represent 1 standard deviation above and below the mean. Subfield Abbreviations include: Cornu Ammonis (CA), Granule Cell and Molecular Layer of Dentate Gyrus (GC-ML-DG); Hippocampus-Amygdala-Transition-Area (HATA); Hippocampal Parcellation (HP).

Figure 3.

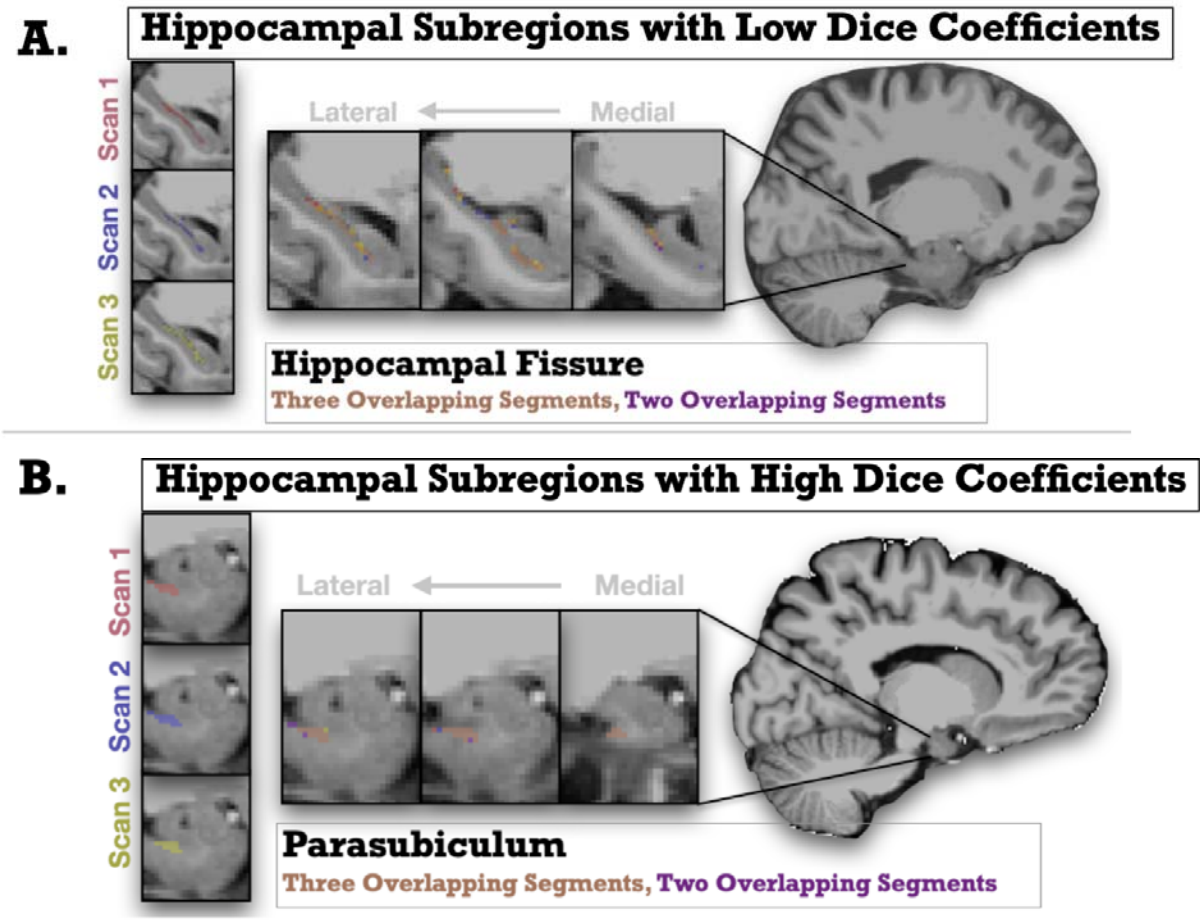

Figure 3: Graphic representations showing magnified depictions of Hippocampal subregions with low and high Dice coefficients (i.e., spatial reliability) from repeated scans (Scan 1 shown in Red; Scan 2 shown in Purple, Scan 3 shown in Yellow). The anatomical (T1w) image underlaid is the unbiased subject template from an example participant. The top panel A represents the hippocampal fissure, an area with low spatial reliability across scans, and the bottom panel B represents the parsubiculum, and area with high spatial reliability. Slices move right to left from medial to lateral.

Figure 4.

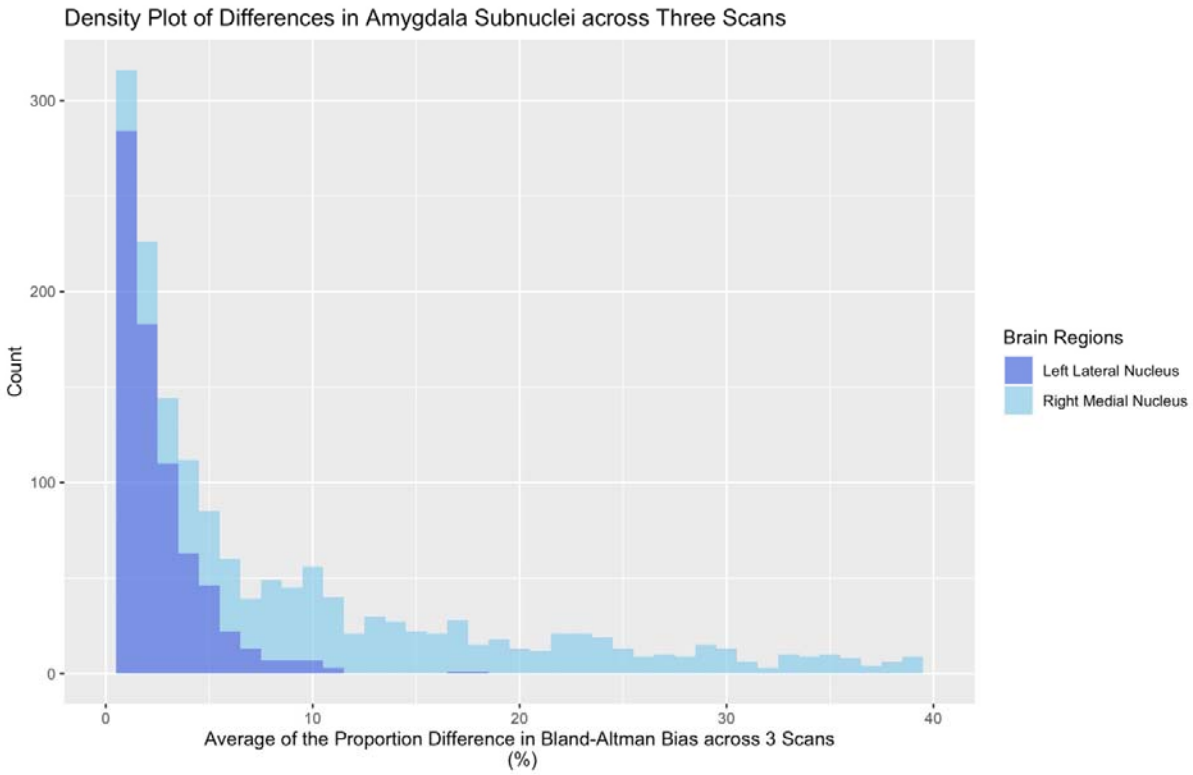

Figure 4: Bland–Altman plots of the average volume difference estimation across Scan 1, Scan 2, and Scan 3 for the left Lateral Nucleus (dark blue) and right Medial Nucleus (light blue). The horizontal axis indicates the average difference in Bland-Altman "*bias*" (difference between subregional volume output for different scans, as a proportion of a region's volume), while the vertical axis indicates the number of scans with a given value. Of note, the left Lateral Nucleus has a low degree of bias (as a proportion of the region's volumes; 0.108%), while the right Medial Nucleus has a fair degree of bias (1.047%).

Figure 5.

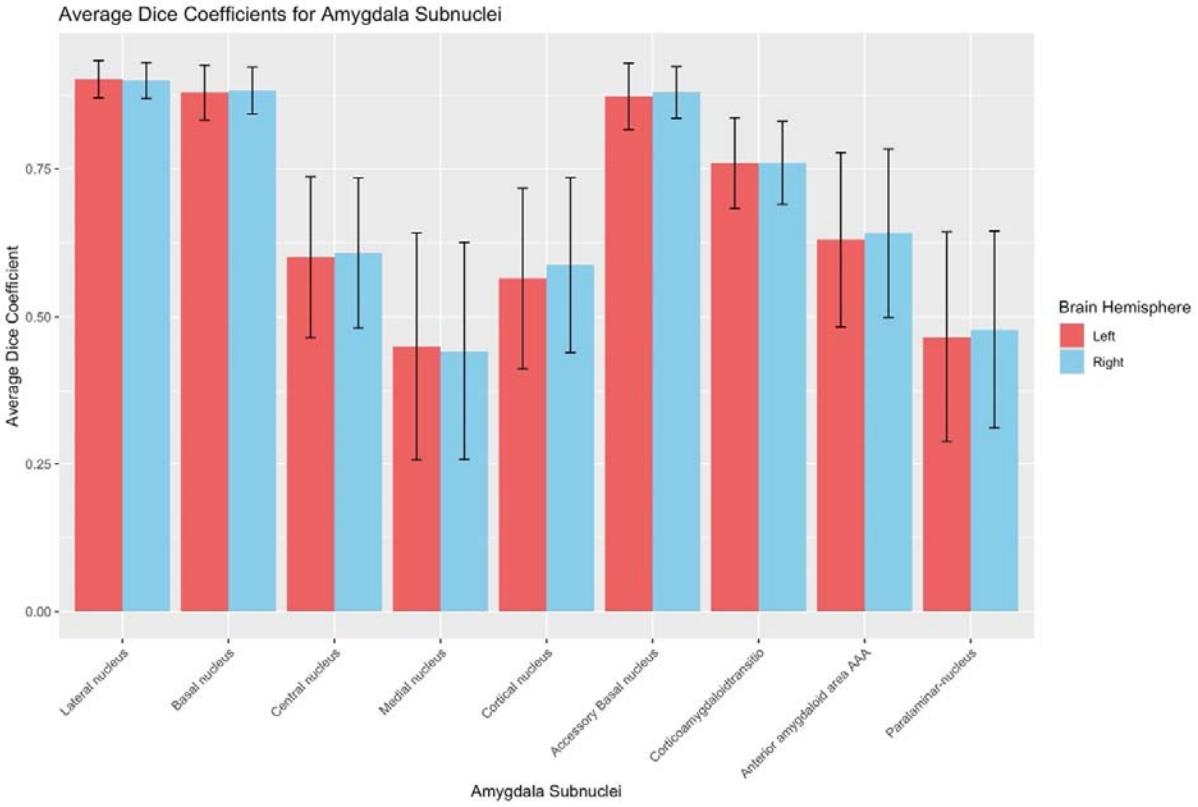

Figure 5. Amygdala Dice Coefficient values for all Amygdala subnuclei. Error bars represent 1 standard deviation above and below the mean.

Figure 6.

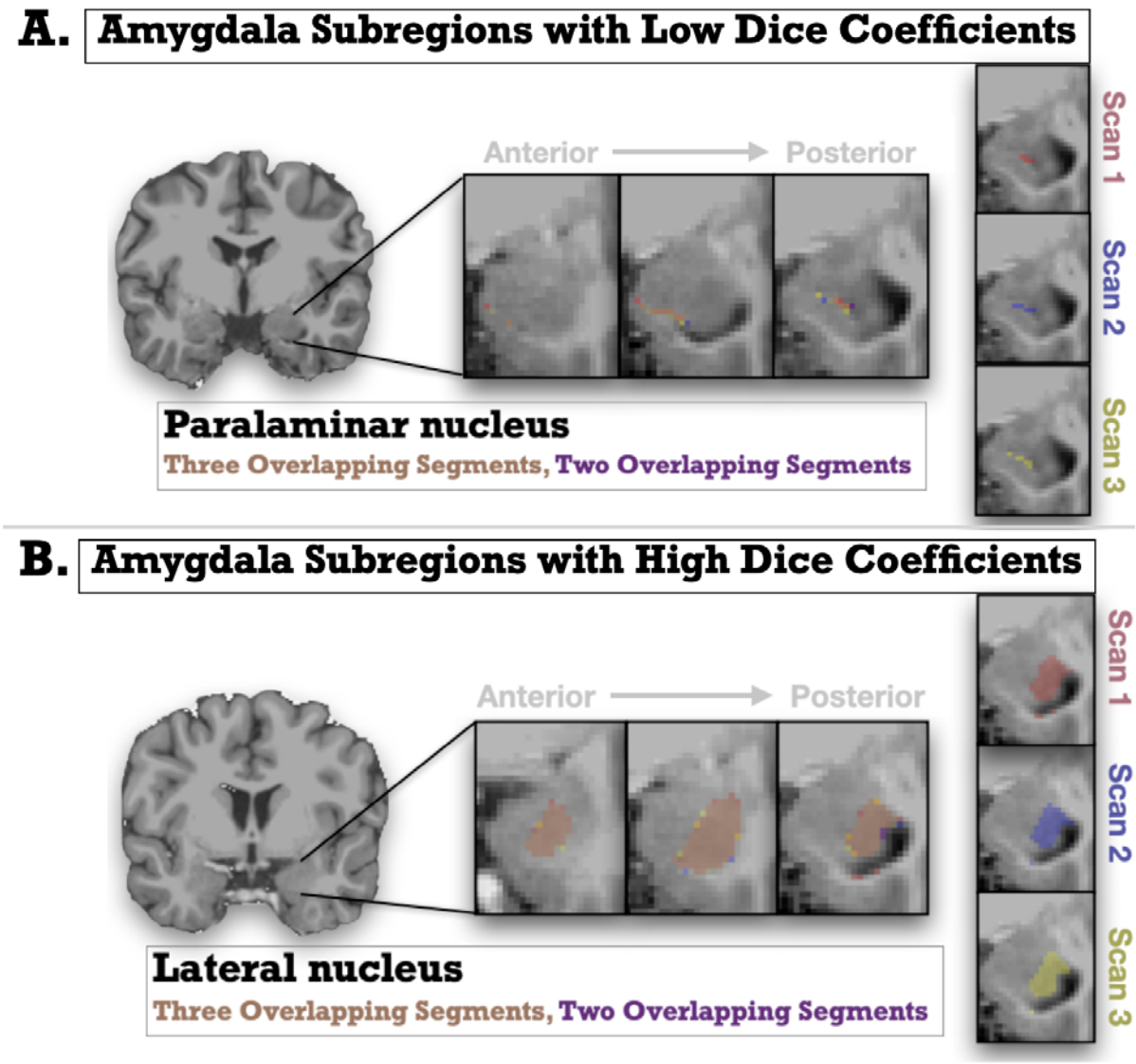

Figure 6: Graphic representations showing magnified depictions of Amygdala subregions with low and high Dice coefficients (i.e., spatial reliability) from repeated scans (Scan 1 shown in Red; Scan 2 shown in Purple, Scan 3 shown in Yellow). The anatomical (T1w) image underlaid is the unbiased subject template from an example participant. The top panel A represents the paralaminar nucleus, an area with low spatial reliability across scans, and the bottom panel B represents the lateral nucleus, and area with high spatial reliability. Slices move right to left from medial to lateral. Multiple slices are depicted left to right, moving anterior to posterior.