Comparison of semi-automated hippocampal subfield segmentation methods in a pediatric sample

Margaret L. Schlichting[1*+], Michael L. Mack[1+],
Katharine F. Guarino[1] & Alison R. Preston[1-3]

[1]Center for Learning and Memory and Departments of [2]Psychology and [3]Neuroscience, The University of Texas at Austin

Running title: Semi-automated hippocampal subfield segmentation in pediatric sample

[+] Authors had equal contributions.

* **Corresponding author**: Margaret L. Schlichting, Ph.D.
Center for Learning & Memory
The University of Texas at Austin
1 University Station, C7000
Austin, TX 78712
megschlichting@gmail.com

ABSTRACT

Episodic memory function has been shown to depend critically on the hippocampus. This region is made up of a number of subfields, which differ in both cytoarchitectural features and functional roles in the mature brain. Recent neuroimaging work in children and adolescents has suggested that these regions may undergo different developmental trajectories—a fact that has important implications for how we think about learning and memory processes in these populations. Despite the growing research interest in hippocampal structure and function at the subfield level in healthy young adults, comparatively fewer studies have been carried out looking at subfield development. One barrier to studying these questions has been that manual segmentation of hippocampal subfields—considered by many to be the “gold standard” approach for defining these regions—is laborious and can be infeasible for large cross-sectional or longitudinal studies of cognitive development. Moreover, manual segmentation requires some subjectivity and is not impervious to bias or error. In a developmental sample of individuals spanning 6-30 years, we compared the performance of two semi-automated segmentation approaches, Advanced Normalization Tools (ANTs) and Automated Segmentation of Hippocampal Subfields (ASHS), to manual subfield delineation on each individual by a single expert rater. Across several quantitative metrics, we found negligible differences in subfield reliability across the child, adolescent, and adult age groups, suggesting that these methods can be reliability applied to developmental studies. We conclude that ASHS outperforms ANTs overall, and is thus preferable for analyses carried out in individual subject space. However, we underscore that ANTs is also acceptable, and may be well-suited for

analyses requiring normalization to a single group template (e.g., voxelwise analyses across a wide age range). Previous work has validated the use of such methods in healthy young adults, as well as several special populations such as older adults and those suffering from mild cognitive impairment. Our results extend these previous findings to show that ASHS and ANTs can also be used in pediatric populations as young as six.

KEYWORDS development; structural MRI; volume; high-resolution MRI; reliability

Recent years have seen increasing research interest in how the hippocampus (HPC) develops, both in terms of structure and function. In particular, work combining high-resolution structural imaging methods with new analysis techniques (Gogtay et al., 2006; Lee et al., 2014; Lin et al., 2013) have suggested that the HPC may continue to change in subtle ways through at least late childhood, and perhaps even into early adulthood. For instance, different developmental trajectories have been observed across the anterior-posterior axis of the hippocampus, with anterior regions generally showing decreases and posterior regions showing increases in size with age (Demaster et al., 2013; Gogtay et al., 2006). Theoretical and animal models also suggest that anatomical pathways within the hippocampal circuit may develop at different rates (Gómez and Edgin, 2015; Lavenex and Banta Lavenex, 2013), which may give rise to the different developmental trajectories sometimes observed across subfields (Tamnes et al., 2014): the *cornu ammonis* (CA) fields, dentate gyrus, and subiculum. These substructures within HPC are known to show their own, unique anatomical and functional characteristics in the mature brain (Manns and Eichenbaum, 2006). Thus, these perspectives suggesting inhomogeneity of development make a host of predictions about when the functions of different substructures—and thus, specific mnemonic behaviors—each reach maturity. Yet, many of these hypotheses remain to be tested empirically in humans.

Answering these questions requires the application of advanced neuroimaging techniques to pediatric populations. For instance, future studies will likely seek to use high-resolution functional MRI (fMRI) to interrogate activation profiles within different subfields of the hippocampus, and characterize how they change over development.

This approach requires not only the acquisition of high-resolution fMRI data, but also the ability to reliably demarcate HPC subfields as anatomical regions of interest (ROIs) and/or normalize individual anatomical images to a custom template generated for the purpose of localizing activations to particular HPC subfields. These methods should be able to be carried out in a consistent manner on participants spanning a wide age range—preferably in a manner that is easily reproducible across studies and research groups.

There is an increasing push toward larger sample sizes in neuroscience research (Button et al., 2013), and developmental work is no exception. If anything, developmental questions may require an even larger sample than would be typical: developmental researchers might investigate individual differences as a function of age, or characterize within-participant development by acquiring data at multiple timepoints in a longitudinal study. As manual delineation of HPC subfields is time-consuming to both master and perform, the large amount of data required by many developmental researchers renders the “gold standard” method intractable. Moreover, while there may be no true replacement for neuroanatomical expertise, even the best and most experienced raters will exhibit some degree of variability, subjectivity, and/or bias (Yushkevich et al., 2015a).

The questions that remain outstanding in the literature—as well as the practical concern of too much data, too little time—motivated us to assess how existing semi-automated HPC subfield segmentation methods compare with a given hand tracing protocol, particularly in younger participants. While at least one published study has used an automated method to investigate the development of hippocampal subfields

(Krogsrud et al., 2014; Tamnes et al., 2014), no quantitative validation of the method was performed. In terms of the HPC overall, one study (Schoemaker et al., 2016) did investigate the correspondence between manual tracing of overall HPC in children and two automated segmentation methods: Freesurfer (Desikan et al., 2006; http://surfer.nmr.mgh.harvard.edu/) and FIRST, part of FSL (Smith et al., 2004; www.fmrib.ox.ac.uk/fsl). Both automated methods failed to reach acceptable levels of reliability; however, it was unclear whether the failure of automated methods was due to the developmental stage of the sample or is a concern about the automated protocols more generally, as similar reliability measures have been reported in adults (Doring et al., 2011; Pardoe et al., 2009). However, other groups have reported success with automated methods for segmenting the overall HPC in preterm neonates (Guo et al., 2015) and toddlers (Gousias et al., 2008). With regards to subfield segmentation, while there is no dearth in literature validating the use of such methods in other special populations such as older adults and individuals suffering from mild cognitive impairment, Alzheimer disease, and psychoses (Pipitone et al., 2014; Yushkevich et al., 2010, 2015b); it is unknown how well automated methods can approximate manually delineated regions in children and adolescents specifically.

In the present study, we sought to assess the validity of two semi-automated subfield segmentation techniques for use in pediatric samples. We chose two methods with which we could use our atlas of choice, allowing us to generate regions that would be directly comparable to our drawn ROIs (**Fig. 1**): reverse normalization of regions drawn on a custom template to each participant's native space using Advanced Normalization Tools (ANTs) (Avants et al., 2011), and automated segmentation on each

participant's brain implemented using the Automated Segmentation of Hippocampal Subfields (ASHS) software (Yushkevich et al., 2015b). Critically, both methods also allowed us to use a single template and analysis strategy across age groups, which—assuming there is no systematic, age-related bias—enable more direct comparisons across groups. We then compared the ROIs generated by each automated method to the manually drawn regions separately for participants in child, adolescent, and adult age groups to test for possible differences in the validity of these approaches across our age range of interest. Results revealed negligible evidence that the reliability of ANTs and ASHS varied as a function of age group. While both methods performed well overall, ASHS outperformed ANTs on several of our metrics. For researchers wishing to employ our segmentation approach on their own datasets, our ANTs template and ASHS atlas have been made available for download here: https://osf.io/hrv9n/.

MATERIALS AND METHODS

Participants

Ninety volunteers participated in the experiment across child (ages 6-11 y; N=31), adolescent (12-17 y; N=25), and adult (18-30 y; N=34) age groups. The consent/assent process was carried out using age-appropriate language in accordance with an experimental protocol approved by the Institutional Review Board at the University of Texas at Austin. For participants under 18, assent was obtained from the participant and permission was obtained from his/her parent or guardian. Adults provided consent.

All participants received monetary compensation and a small prize for their involvement in the study.

Participants were screened for psychiatric conditions using the Child Behavior Checklist (CBCL; completed by the parent/guardian of participants aged 6-17; Achenbach, 1991) and the Symptom Checklist 90-Revised (SCL-90-R; adults; Derrogatis, 1977). IQ was assessed using the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II; Wechsler, 1999). The intelligence measure of interest was the full-scale IQ composite score (FSIQ-2), which includes vocabulary and matrix reasoning subtests.

From the original group of 90 participants, individuals were excluded from subsequent analysis if they met any of the following criteria: (1) CBCL score in the clinical range (N=1 child; N=1 adolescent) or SCL-90-R score greater than 1 SD above the mean of a normative sample (N=9 adults); (2) presence of a psychiatric condition (N=1 adult); (3) did not complete the MRI portion (N=3 children; N=4 adults); (4) handedness concerns (N=1 adolescent); (5) MRI data not of acceptable quality (N=4 children); or (6) automated segmentation failed (N=1 adolescent; N=2 adults). Participants were excluded for automated segmentation failure when visual inspection revealed that the hippocampus as a whole was grossly mislabeled (e.g., with HPC extending through white matter and into medial temporal lobe cortex). For two of these participants (adults), both ASHS and ANTs failed; only ANTs failed for the third participant (adolescent). No participants scored below our inclusion threshold for IQ (> 2 SD below the mean). The final sample included a total of 64 right-handed participants (23 children, 13 females, ages: 6.08-11.83 y, 9.38 ± 0.39 y, FSIQ-2: 84-142, 119.52 ±

2.85; 22 adolescents, 10 females, ages: 12.08-17.33 y, 14.23 ± 0.35 y, FSIQ-2: 92-130, 111.27 ± 2.48; 19 adults, 10 females, ages: 18.67-28.92 y, 23.86 ± 0.78 y, FSIQ-2: 92-129, 113.00 ± 2.76).

Experiment Overview

The experiment comprised two visits. On the first visit, participants were exposed to the MRI environment using a mock scanner, completed paper-based screening measures (CBCL or SCL-90-R; WASI-II), and performed a battery of cognitive tasks (not discussed here). MRI scanning took place during the second visit.

MR Data Acquisition

Imaging data were acquired on a 3.0T Siemens Skyra MRI. Two to three oblique coronal T2-weighted structural images were acquired perpendicular to the main axis of the HPC (TR=13150 ms, TE=82 ms, 512 x 60 x 512 matrix, 0.4 x 0.4 mm in-plane resolution, 1.5 mm thru-plane resolution, 60 slices, no gap). Coronal images of acceptable quality as determined by visual inspection (e.g., absence of motion artifacts that would prevent visualization of the hippocampal sulcus) were coregistered using ANTs (Avants et al., 2011) and averaged to boost SNR, yielding a single mean coronal image per participant. A T1-weighted 3D MPRAGE volume (256 x 256 x 192 matrix, 1 mm^3 voxels) was also collected.

Baseline method: Manual hippocampal subfield delineation

We used manual demarcation to define subfield on each participant's anatomical image (hereafter denoted Manual), which is typically considered the "gold standard" of anatomical volume assessment (Rodionov et al., 2009). As such, we treated these segmentations as the ground truth; all semi-automated methods described were compared to Manual in a pairwise fashion. HPC regions of interest (ROIs) were delineated on each participant's mean coronal image by a single rater (KFG) using established guidelines (West and Gundersen, 1990). We chose to have a single rater perform the segmentations in order to reduce concerns of inter-rater reliability, a majority source of variance associated with manual tracing. The rater was blind to participant identity (including but not limited to age, sex, and performance on behavioral tasks), and images were cropped to obscure overall head size.

HPC was segmented into the following subfields: *cornu ammonis* fields 1 (CA_1) and 2/3 (combined; $CA_{2/3}$), dentate gyrus (DG), and subiculum. Segmentation was performed across the entire extent of the HPC long axis, with the exception of the most posterior slices on which subfields could not be reliably delineated. For this region, we created a combined posterior HPC ROI. All subfields and posterior HPC were summed to create overall HPC ROIs. As functional neuroimaging studies often interrogate activation within a combined DG/$CA_{2/3}$ region, we also summed individual DG and $CA_{2/3}$ regions to create DG/$CA_{2/3}$.

We assessed both intra- and inter-rater reliability to validate our manual segmentation approach using intra-class correlation (ICC). These results are reported in detail elsewhere (Schlichting et al., under review). Briefly, our results were in the same range as previous reports for manual demarcation of hippocampal subfields (Lee et al., 2014; Mueller et al., 2007; Wisse et al., 2012; Yushkevich et al., 2010), suggesting that manual segmentation was reliable across time and across raters.

Automated methods for comparison

Comparison method 1: Custom template ROIs reverse normalized (ANTs). For the first comparison method (**Fig. 1A**), we defined ROIs on a template brain image, and then back-projected the regions into each participant's native space. We first generated a series of custom templates from the mean coronal images of a subset of participants with canonical hippocampi. Template generation and normalization was carried out using ANTs version 2.1.0 (Avants et al., 2011). Multiple templates were generated using different subsets of participants. We then selected the best group template, i.e., one that was free from artifacts and for which the HPC subfield landmarks could be visualized. Efforts were made to include participants who spanned our age range in the generation of all templates. The final chosen template used for both comparison methods 1 and 2 was created from the mean coronal images of 10 individuals (age range: 7.5-28.42 y; N=3 children, N=3 adolescents, N=4 adults).

The manual rater (KFG) then segmented the HPC on the group template into subfields using the protocol described above for the manual ROI delineation. In other

words, the subfields were demarcated on the group coronal template in the same way as they were for an individual brain. We then back-projected all ROIs into the native space of each participant as follows. First, we computed the nonlinear transformation from the individual's mean coronal to the group template using the following settings: image metric: probabilistic; transformation: symmetric normalization; regularization: Gaussian. We then applied the inverse transformation to each ROI.

This method does require some time and expertise on the part of the researcher to implement. First, it may be the case that multiple templates need to be generated to yield one that will allow for manual segmentation and satisfactory normalization. Second, the researcher must manually segment the regions on the template according to their desired protocol. Thus, this procedure is not fully automated; however, it does significantly cut down the hands-on time required by the researcher relative to manual segmentation of each individual, making it tractable for larger N studies.

Comparison method 2: Custom template ROIs reverse normalized using ROI-guided methods ($ANTs_{ROI}$). The second comparison method used a procedure identical to the one described for ANTs above, with the single exception that regions were back-projected into native space using nonlinear transformations computed with ROI-guided methods implemented using landmarkmatch, part of the ANTs toolbox (**Fig. 1A**). Hereafter, this method of ROI-guided ANTs normalization is denoted $ANTs_{ROI}$. Specifically, the Manual HPC for each individual was used to guide the normalization to the custom group template, with the Manual HPC template ROI serving as the target

(weight: 1). The inverse of the computed transformation was then applied to the group template ROIs, such that regions were back projected to native space.

As with ANTs, $ANTs_{ROI}$ requires template selection and manual demarcation on the part of the researcher. In addition, a significant downside to $ANTs_{ROI}$ as compared with ANTs is that implementing this approach requires a rater to trace HPC on each participant's brain individually. This practical concern should be taken into consideration when choosing a segmentation method.

Comparison methods 3 and 4: ASHS automated segmentation using custom atlases. For the third and fourth comparison methods (both depicted in **Fig. 1B**), we built custom atlases for use with the Automated Segmentation of Hippocampal Subfields (ASHS) software (version 0.1.0, rev 103 downloaded on 4/14/2016) (Yushkevich et al., 2015b). ASHS is an open-source software package that uses both T1- and T2-weighted images to automatically segment the medial temporal lobe into subregions. It can be used with an included atlas, or retrained to use any segmentation protocol chosen by the user. For the present study, we generated custom atlases based on the Manual ROIs for a subset of participants with canonical hippocampi. Nine participants from each of the three age groups were selected for atlas generation (age range: 6.08-28.75 y). We then built two atlases: one that combined across DG and $CA_{2/3}$ (ASHS combined, hereafter $ASHS_C$), and one that included them as separate regions (ASHS separated, hereafter $ASHS_S$). These two comparison methods were otherwise identical. Following atlas generation, automated segmentation was carried out on all participants.

Manually delineated ROIs are needed for the initial step of generating a custom atlas. However, once an atlas has been created, it can easily be applied to an entirely new sample, which would make the segmentation entirely automated. Users wishing for a fully automated pipeline may also download and use an existing, freely available atlas.

Volume extraction

Each comparison method resulted in HPC, CA_1, $CA_{2/3}$, DG, $DG/CA_{2/3}$, and subiculum regions in each participant's native space. Because the most posterior portion of the hippocampus was not segmented into subfields on the majority of participants, we consider reliability of subfield delineation only on slices for which subfields were drawn. Raw volumes (not adjusted for differences in intracranial volume) were extracted for all methods, and the reliability of the automated methods were compared across age groups using various metrics as described below. Basic quality assurance was performed to exclude participants (N=3) whose overall HPC was grossly mislabeled, as described above. This was the only quality control we carried out for the automated segmentations; no manual editing of regions was performed.

Comparing automated segmentation methods with manual demarcation

Spatial overlap (DSC). We first interrogated how much each automated method agreed with the Manual regions in terms of their spatial overlap. Spatial overlap was

indexed using Dice similarity coefficients (DSC; Dice, 1945). For region X segmented using methods A and B, DSC is defined as:

$$DSC_{X(A,B)} = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)}$$

where A and B are the set of voxels marked as region X by methods A and B, respectively. DSC was computed for each participant, ROI, and comparison method; the resulting values were then averaged across participants within each age group.

Edge agreement across methods. While DSC quantifies the amount of overall spatial overlap between two methods, it does indicate where the disagreements among methods lie. To further characterize the spatial locations at which our automated methods agree (or disagree) with our Manual subregions, we indexed agreement at particular voxels in a mapwise fashion (see Yushkevich et al., 2015 for a similar approach). For each participant and method, we created an image indicating edge voxels (i.e., voxels on either side of the boundary between two subfields, or between a subfield and the outside of HPC). We then compared the edges for each method to the Manual edge map, characterizing agreement as voxels that were marked as edges by both methods. These maps were then normalized to the group template for display and averaged across individuals within each age group to serve as a qualitative representation of edge agreement. The intensities of the resulting maps thus indicate the proportion of participants that showed agreement between the two methods at that voxel.

Volume correspondence: Intra-class correlation. We next assessed the degree of correspondence between regional volumes determined using each method, irrespective of spatial agreement. We computed intra-class correlations (ICC) to measure consistency of averaged measures using a two-way random effects ANOVA model. ICC values near 1 indicate highly consistent segmentation, while values near 0.5 indicate unreliable segmentation. Prior work on hippocampal subfields using the same ICC measure has considered values exceeding 0.70 to represent good consistency (Mueller et al., 2007).

Bias. One possibility is that the degree to which volumes are over- or under-estimated by a given method varies as a function of the region's size. This might be of particular concern for researchers investigating hippocampal subfields (as subfields vary substantially in size) and/or their development (as overall size will differ across age groups). To assess this possibility, we generated Bland-Altman plots (Bland and Altman, 2007), which show the difference in volume of the two methods (e.g., ANTs-Manual on the y-axis) as a function of the average of the two methods (x-axis). If the automated method is not systematically over- or under-estimating the region volumes relative to Manual, the difference values should be centered around zero. Additionally, we performed ANCOVAs to determine whether there were significant main effects of average volume, age, and/or a volume x age interaction. Significant effects of volume indicate bias in the volume estimation across all three age groups, with the size of the ROI predicting the degree to which volume is over or underestimated by the automated method. Significant effects of age indicate that the degree of volume under- or

overestimation differs across the three age groups. A significant volume x age interaction would indicate that the bias profile differs across the age groups.

Assessing reliability within segmentation method

Volume correspondence within an individual: inter-hemispheric correlations. An individual's anatomy is generally relatively similar across hemispheres (Allen et al., 2002). As such, one metric that has been proposed to capture reliability within method (Schoemaker et al., 2016) quantifies the degree to which a given structure's size in one hemisphere predicts its size in the other. We performed across-participant Pearson's correlations of left and right hemisphere volumes for each region (hereafter termed inter-hemisphere correlations, IHC).

Statistical analyses

For all quantitative metrics (DSC, ICC, IHC, and Bias), we used a nonparametric approach to compute 95% confidence intervals and p values. We resampled participants with replacement across 1,000 iterations, each time computing the statistic(s) of interest. Inferences were made on the basis of these bootstrapped distributions. All reported statistics were derived using this approach.

For DSC, ICC, and IHC measures, child and adolescent groups were each compared in a pairwise fashion to the adult group, which was treated as the baseline. Specifically, we generated a bootstrapped distribution of the difference between the two

groups, which we then compared to zero to generate p values. In the case of the Bias analysis, F statistics for the main effects and interaction were computed across iterations, and the bootstrapped distribution was compared to one.

RESULTS

Comparing automated segmentation methods with manual demarcation

Our primary goal was to determine how well each automated method corresponded with those regions delineated by an expert human rater. We thus assessed spatial overlap, edge agreement, volume correspondence, and potential bias of each automated method—ANTs, $ANTs_{ROI}$, $ASHS_S$, and $ASHS_C$—relative to Manual.

Spatial overlap (DSC). The results of the spatial overlap analysis are presented in **Table 1** and **Figure 2**. DSC values neared or exceeded the agreement typically achieved by two human raters (ranging roughly from 0.7-0.85 in Olsen et al., 2013; though the specific values will be resolution-dependent, Yushkevich et al., 2015b) for all regions and methods with the exception of $CA_{2,3}$ using ANTs and $ANTs_{ROI}$. For all regions, the ASHS approaches yielded better reliability than the ANTs methods; comparing the two ANTs approaches, $ANTs_{ROI}$ outperformed ANTs in all cases, suggesting that landmark-guided normalization may be superior when aiming to achieve maximal spatial overlap. Encouragingly, there was limited evidence that age impacts spatial overlap in the methods we tested: children and adolescents showed a small but

reliable difference in DSC compared with adults only for the overall HPC when using $ANTs_{ROI}$.

Edge agreement across methods. A voxelwise depiction of edge agreement is displayed in **Figure 3**. The boundaries between CA_1 and DG, which largely rely on boundaries visible on MRI, were generally quite high across most of the anterior-posterior extent of the HPC, as well as across age groups. The outer edge of the overall HPC was also quite reliable across all methods, particularly on the lateral portions of the structure; there was relatively less agreement on how far DG/$CA_{2,3}$ (particularly in the HPC head) and subiculum should extend in the medial and ventral directions, respectively. As would be expected, agreement was lower for those divisions lacking visible anatomical boundaries, such as between CA_1 and subiculum and between $CA_{2,3}$ and DG in the head. It did seem to be the case that there was less overall agreement about where edges should be placed in the head relative to the body of the HPC. Despite these differences in agreement across the structure, there were no apparent differences in this pattern across age groups.

Volume correspondence: Intra-class correlation. In the next analysis, we investigated the degree to which each subfield volumes derived from the semi-automated methods corresponded with Manual, irrespective of how their spatial agreement. These results are shown in **Table 2** and **Figure 4**. Comparing ICC measures, we found that all methods provided acceptable correspondence for overall HPC, CA_1, and subiculum. Reliability was lower, and more variable, in $CA_{2,3}$, DG, and

the combined DG/$CA_{2,3}$ region. It is worth noting that the ASHS methods yielded higher ICC values than either ANTs method almost universally for these regions. Echoing the DSC results, age-related differences were limited; there was significantly better volume agreement in the child group relative to the adults for just $CA_{2,3}$ and DG/$CA_{2,3}$ when using $ANTs_{ROI}$. Thus, using ASHS (particularly with larger regions, or when combining DG/$CA_{2,3}$), the researcher can be reasonably confident that the reliability measures would easily exceed proposed thresholds for good consistency (ICC > 0.7; Mueller et al., 2007) for all regions in children through adults.

Bias. We generated Bland-Altman plots to look for potential bias in the volumes estimated by each automated method, shown in **Figures 5-8**. These plots show each automated method-Manual segmentation difference as a function of the average volume estimated across the two methods being compared. First, none of the automated methods showed evidence of overall bias; that is, the points are all centered around a difference of zero on the y-axis, meaning that none of the methods systematically over- or under-estimate the volumes relative to Manual. Second, we performed a series of ANCOVAs to test how the average volume, the age group, or the volume x age group interaction impacted the degree of discrepancy we observed between the two methods (**Table 3**). The majority of regions (subiculum being the notable exception) showed a significant main effect of volume for all four methods; this indicates that the larger a region is, the more its volume tends to be underestimated by the automated methods. Significant main effects of age were found for overall HPC using $ANTs_{ROI}$, $ASHS_S$, and $ASHS_C$; CA_1 using ANTs and $ANTs_{ROI}$; and $CA_{2,3}$ using

$ASHS_S$. These results suggest that the degree of bias differs reliably between the three groups—although, notably, the group means were near zero in all cases, suggesting that these small differences in bias may still leave all groups within the acceptable range. Only one region showed a volume x age group interaction: subiculum when using $ASHS_S$. Taken together, these results suggest that while present in some cases, the effects of age group on the degree of bias observed pale in comparison to the effect of region size.

Assessing reliability within segmentation method

Volume correspondence within an individual: inter-hemispheric correlations. In addition to characterizing how well each automated method corresponds with the Manual segmentation, we also wanted to quantify how consistent each method is with itself across multiple hippocampi exemplars. Previous work (Schoemaker et al., 2016) has investigated the correlation between volumes in the left and right hemispheres as an index of this within-method agreement, since individuals tend to have relatively similar structural volumes across hemispheres. We computed this measure for all methods including Manual, which serves as a baseline.

These data are displayed in **Table 4** and **Figure 9**. First, all regions showed a significant relationship between left and right hemispheres across all methods, with the exception of $CA_{2,3}$ using the automated approaches (particularly in the adult group). For the Manual ROIs, we observed no significant age group effects, indicating that the degree of correspondence across hemispheres changes little across development.

However, when using the automated approaches, several regions did show significantly greater IHC for the child and/or adolescent groups relative to the adults: $CA_{2,3}$ for ANTs and $ASHS_S$, $DG/CA_{2,3}$ for both ANTs and $ANTs_{ROI}$, and DG for $ANTs_{ROI}$. Adolescents showed significantly lower IHC relative to the adults for subiculum using both $ASHS_S$ and $ASHS_C$, as well as $DG/CA_{2,3}$ for $ASHS_C$. It seems to be the case that many of these differences that emerged for the automated methods were driven by under- or overestimating the level of symmetry for a single group relative to the Manual regions, rather than a fundamental change in the pattern shown across all three groups. For instance, in subiculum, three of the four automated methods seem to result in increased IHC relative to Manual in the adult group, while the estimates for the other groups appear to remain unchanged.

DISCUSSION

We compared HPC subfield regions of interest defined using two semi-automated methodological approaches to those demarcated by an expert human rater. Participants were children, adolescents, and adults, spanning an age range of approximately 6 through 30 years. Note that here, we consider only the HPC proper, and did not investigate subregions of MTL cortex; however, there is no reason why, in principle, similar methods could not be used to segment cortical regions as well. The validation of these methods for MTL cortical segmentation in pediatric samples remains an open question for future studies.

We chose methods that allowed us the flexibility to match the tracing protocol to the one we implemented manually, removing the concern that differences in the atlases used across methods may somewhat artificially lower performance (de Flores et al., 2015; Schoemaker et al., 2016; Wisse et al., 2014). We found generally good reliability across the methods, with markedly few differences in agreement for children and adolescents relative to the adult comparison group. These results suggest that ANTs and ASHS approaches may be used by researchers without fear of introducing unequal measurement error across ages, with a few caveats discussed in detail below. Our findings extend previous work assessing the reliability of automated methods designed to segment overall HPC in pediatric samples (Guo et al., 2015; Schoemaker et al., 2016). While these prior studies have yielded mixed results, our data suggest that there do exist methods that will segment the HPC into subfields (in addition to overall HPC) in a way that approximates manual segmentation by an expert human rater.

One subfield that stands out as an exception to the overall high reliability we found is $CA_{2,3}$, which shows the lowest reliability systematically across our measures. $CA_{2,3}$ is an extremely small and slender region, which makes it difficult to achieve high levels of spatial overlap and volume correspondence. Agreement for this region, even between expert human raters, is often markedly lower than for other subfields (Yushkevich et al., 2010, 2008). Our findings suggest that the best approach for researchers seeking to isolate $CA_{2,3}$ is to use ASHS, which yielded the highest levels of reliability; agreement was similar to published levels of inter-rater reliability (Wisse et al., 2012) using several of our metrics (e.g., nearing or exceeding 0.7 for all groups in DSC). However, as a cautionary note, $CA_{2,3}$ did show strong and significant evidence for a

relationship between structure size and the degree of bias of all the subfields, and it was present across all methods tested here. This finding indicates that these automated methods will underestimate volume for individuals with larger $CA_{2,3}$ regions, which may reduce the ability to detect population differences related to typical development, aging, or disease (de Flores et al., 2015). On the upside, however, $CA_{2,3}$ did not show any evidence for significant age-related reliability differences in the present study. Researchers choosing to investigate $CA_{2,3}$ should keep these limitations in mind, and consider collapsing across regions if allowable in the context of their research questions. If combining across regions is not possible, ASHS appears better suited than ANTs for approximating manual subfield segmentation of $CA_{2,3}$.

One interesting pattern that stands out in our IHC analysis is the trend for a general decrease in symmetry over the course of development, a finding particularly evident in the subfields. Lateralized effects are often observed in terms of HPC task-based activation (Addis et al., 2011; Glosser et al., 1995; Golby et al., 2001; Kelley et al., 1998; Mack and Preston, 2016; Martin et al., 1997; Schlichting et al., 2014; Zeithamova and Preston, 2010), suggesting that the two hippocampi may serve unique and specialized roles. It may be the case that the general decrease in symmetry we observed is related to emerging specialization, as the two hemispheres progress along distinct developmental trajectories. Consistent with this notion, prior work has shown greater developmental change in the right than left HPC overall (Dennison et al., 2013). Similar ideas have been proposed for developmental emergence of hemispheric specialization in the visual stream (Behrmann and Plaut, 2015; Dundas et al., 2013);

whether developmental asymmetry similarly reflects shifts in functional specialization across hemispheres in HPC subfields as well remains to be tested in future studies.

Strengths of automated methods

Previous reports have used automated methods to draw conclusions about HPC structure in pediatric populations (Guo et al., 2015; Krogsrud et al., 2014; Lin et al., 2013; Tamnes et al., 2014). For instance, one study (Lin et al., 2013) used automated methods to characterize maturational differences in HPC shape among children from 6-10 years of age. Their approach yielded findings that were not detectable when comparing volume alone, demonstrating the utility of applying such analyses to developmental questions. The authors applied three different automated methods and found that they gave similar results, suggesting the reliability of these methods even in their young participants. However, to the best of our knowledge, only two reports have directly tested the validity of automated segmentation methods in developmental samples by comparing them to manual tracing. One such study (Schoemaker et al., 2016) found that both Freesurfer and FSL-FIRST overestimated overall HPC volumes, suggesting that these methods are not good approximations of manual tracing. Importantly, however, this problem does not seem to be unique to the children that were participants in that study, as other papers in adults have shown similarly low reliability levels using these methods (Doring et al., 2011; Pardoe et al., 2009). Another report used MAGeT-Brain to segment the hippocampi of preterm neonates (Guo et al., 2015). That paper showed that segmentations derived from the automated methods were comparable to those performed manually, suggesting that automated methods can (in

theory) be used with even the youngest of participants. However, these mixed findings underscore that the validity of the particular automated method being used must be assessed, as no two automated methods are the same in all cases.

Our findings extend beyond prior work to show that there exist automated methods for HPC subfield segmentation that can be used reliably in pediatric populations. These results are welcome news for several reasons. While manual tracing is considered the “gold standard” approach for subfield segmentation, it has undeniable drawbacks: it is a laborious, time-consuming process that requires extensive training and anatomical expertise; it requires subjective decisions to be made on the part of the rater, leading to variability both within and across individuals—even those trained on the same tracing protocol—and is susceptible to bias and error; and varies substantially across research groups (Yushkevich et al., 2015a). These issues are each expounded upon in turn below.

Both automated methods described in this paper require relatively little hands-on involvement on the part of the researcher. For a highly experienced rater at the resolution described in this paper, manual delineation of subfields takes roughly two hours; for less experienced raters, the process will take much longer. Automated methods require significantly less attention: once a template or atlas is generated, the ANTs approach we describe requires manual delineation on just one brain (the template), representing an investment of just a few hours; for ASHS, virtually no hands-on researcher time is needed. Depending on the resources available, these methods may require several hours of computation time to complete for a single participant; however, this is time during which the researcher can be doing other things. The

knowledge of HPC subfield anatomy required is also quite limited. There are options available—such as using ASHS with an existing atlas—that require little anatomical expertise to carry out, and (in our experience) have a very low failure rate.

A clear strength of automated approaches is that they reduce the inconsistencies associated with human raters (Lerma-Usabiaga et al., 2016). Such inconsistencies may arise from simple differences in subjective tracing decisions across individuals (e.g., one rater tending to draw regions slightly more “generously” than another) or even within an individual, across occasions. Most relevant to the present report, knowledge of participant age might subconsciously influence raters to delineate regions in a certain way. While measures can be taken to avoid this issue in the case of development (e.g., by blinding the rater to subject identity and cropping the field of view to obscure identifying features like head size), this approach might be difficult or impossible in other special populations (e.g., those with brain damage or obvious atrophy). Thus, automated approaches that remove human subjectivity may be an ideal solution. Unlike other automated methods (Desikan et al., 2006), those we describe here—particularly $ANTs_{ROI}$ and ASHS—do not require manual editing or painstaking examination of regions. While basic quality assessment should certainly be performed, failures with these methods tend to be extreme (though rare) and quite obvious, even to an untrained eye. This fact reduces not only the burden of anatomical knowledge required, but also alleviates issues associated with the potential bias introduced by human subjectivity, even compared with other automated methods.

The way manual tracing is implemented also varies substantially across research groups (Yushkevich et al., 2015a), making it difficult to compare findings across studies.

Both ANTs templates and ASHS atlases are small enough in terms of file size that they can be easily shared with other research groups and/or posted on a webpage to be made freely available for download. These methods can be readily applied to new datasets, thus enabling direct comparisons across studies—even those carried out by different research groups.

In discussing these features, it is clear that automated segmentation methods align well with the aspirations expressed by so many in the field (e.g., Gorgolewski and Poldrack, 2016): science should be more reproducible and open. Moreover, in efforts to make findings more replicable in future studies, larger sample sizes are becoming the goal for some and the norm for many, particularly when individual differences are of interest. This makes manual delineation of hippocampal subfields an intractable approach. In developmental work specifically, this problem is compounded when individual differences vary by age, and/or when data is collected from the same large number of participants over multiple years. The automated methods described here would make hypotheses falling into these categories addressable in a way that is not only logistically feasible but also free from the bias and subjectivity introduced by a human rater.

Suggestions for the field

We believe that the automated methods discussed here may be even better suited than manual tracing to answer the kinds of questions posed by developmental researchers, given the considerations about subjectivity and bias described above. We thus present the following recommendations for developmental researchers interested

in using automated approaches to study HPC subfield anatomy or function. These ideas are described with the caveat that we did not examine all possible automated segmentation methods; recommendations are based on the comparison of just those we did test.

When possible, analyses should be done in each participant's native space using ROIs produced by ASHS. This procedure could include not only analysis of volume but also function, with these subfields serving as anatomical ROIs that can be used to investigate functional engagement or as masks for multivariate searchlight or decoding analyses. We also encourage researchers to consider these subfields split into anterior-posterior head, body, and tail segments, as these regions seem to both support different functions in adults and also show different developmental trajectories (Moser and Moser, 1998; Poppenk et al., 2013; Strange et al., 2014).

When voxelwise analyses are needed, our findings suggest that generating a custom coronal template—importantly, including participants representative of the whole age range—is a valid approach. While the use of the MNI template brain has been previously validated for use with children as young as 7 (Burgund et al., 2002; Kang et al., 2003), this template is not appropriate for localizing activations to specific hippocampal subfields. Our study is the first that we know of to generate a single custom group template spanning a developmental sample of the kind that would be used in high-resolution imaging studies optimized for the medial temporal lobes. This single group template is an important feature of the present method, as it enables direct (i.e., voxelwise) comparisons across individuals of different ages. We found limited age-related differences in our metrics for both ANTs methods, suggesting that the ability to

normalize to the group template does not substantially differ for children and adolescents relative to adults. We have a slight preference for using $ANTs_{ROI}$ over ANTs in this kind of analysis strategy, given that spatial overlap—the metric that, we reason, is most directly related to the validity of warping each subject's functional data or statistical maps to the group template—as superior for the ROI-guided implementation of ANTs across all of the regions we examined. Notably, researchers could opt to use the overall HPC derived from ASHS rather than manually trace the region on each participant to use the ROI-guided normalization, substantially reducing the time burden associated with the $ANTs_{ROI}$ method.

Conclusions

Recent years have seen an increase in the number of studies on HPC subfield function. It is additionally becoming clear that many interesting questions relating to hippocampal development remain to be answered—not only in terms of characterizing the typical developmental trajectory, but also to better understand how the structural development goes awry in less typical scenarios like significant early life stress (Teicher et al., 2003), childhood obesity (Chaddock et al., 2010), and neurodevelopmental disorders (Schumann et al., 2004). Our findings suggest that automated subfield segmentation techniques can be applied to healthy individuals ranging in age from 6-30. While the validity of applying these methods to atypically developing samples remains to be explored, the present results show the promise in this avenue. The ability to readily apply such methods to a diverse sample may result in increased sensitivity for diagnosis, making these findings of relevance to basic researchers and clinicians alike.

Funding

This work was supported by the National Institutes of Health (grant numbers R01MH100121 and R21HD083785 to ARP and F32MH100904 to MLM); the National Science Foundation CAREER Award (grant number 056019 to ARP); a University of Texas Research Grant to ARP; and the Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program (MLS). The authors declare no competing financial interests.

Acknowledgments

Many thanks to Jessica Church-Lang, Tammy Tran, and Amelia Wattenberger for assistance with participant recruitment, data collection, and helpful discussions. We also thank the Texas Advanced Computing Center (http://www.tacc.utexas.edu) at The University of Texas at Austin for providing critical computing resources.

REFERENCES

Achenbach, T.M., 1991. Manual for the Child Behavior Checklist/4-18 and 1991 profile. Department of Psychiatry, University of Vermont, Burlington, VT.

Addis, D.R., Cheng, T., Roberts, R.P., Schacter, D.L., 2011. Hippocampal contributions to the episodic simulation of specific and general future events. Hippocampus 21, 1045–52. doi:10.1002/hipo.20870

Allen, J.S., Damasio, H., Grabowski, T.J., 2002. Normal neuroanatomical variation in the human brain: An MRI-volumetric study. Am. J. Phys. Anthropol. 118, 341–358. doi:10.1002/ajpa.10092

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–44. doi:10.1016/j.neuroimage.2010.09.025

Behrmann, M., Plaut, D.C., 2015. A vision of graded hemispheric specialization. Ann. N. Y. Acad. Sci. 1359, 30–46. doi:10.1111/nyas.12833

Bland, J., Altman, D., 2007. Agreement between methods of measurement with multiple observations per individual. J. Pharm. Stat. 17, 571–582. doi:10.1080/10543400701329422

Burgund, E.D., Kang, H.C., Kelly, J.E., Buckner, R.L., Snyder, A.Z., Petersen, S.E., Schlaggar, B.L., 2002. The feasibility of a common stereotactic space for children and adults in fMRI studies of development. Neuroimage 17, 184–200. doi:10.1006/nimg.2002.1174

Button, K.S., Ioannidis, J.P. a, Mokrysz, C., Nosek, B. a, Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–76. doi:10.1038/nrn3475

Chaddock, L., Erickson, K.I., Prakash, R.S., Kim, J.S., Voss, M.W., Vanpatter, M., Pontifex, M.B., Raine, L.B., Konkel, A., Hillman, C.H., Cohen, N.J., Kramer, A.F., 2010. A neuroimaging investigation of the association between aerobic fitness, hippocampal volume, and memory performance in preadolescent children. Brain Res. 1358, 172–183. doi:10.1016/j.brainres.2010.08.049

de Flores, R., La Joie, R., Landeau, B., Perrotin, A., Mézenge, F., de La Sayette, V., Eustache, F., Desgranges, B., Chételat, G., 2015. Effects of age and Alzheimer's disease on hippocampal subfields. Hum. Brain Mapp. 36, 463–474. doi:10.1002/hbm.22640

Demaster, D.M., Pathman, T., Lee, J.K., Ghetti, S., 2013. Structural development of the hippocampus and episodic memory: developmental differences along the anterior/posterior axis. Cereb. Cortex 24, 3036–3045.

Dennison, M., Whittle, S., Yücel, M., Vijayakumar, N., Kline, A., Simmons, J., Allen, N.B., 2013. Mapping subcortical brain maturation during adolescence: Evidence of

hemisphere- and sex-specific longitudinal changes. Dev. Sci. 16, 772–791. doi:10.1111/desc.12057

Derogatis, L.R., 1977. SCL-90-R: Administration, scoring and procedures: Manual 1. Clinical Psychometric Research, Baltimore, MD.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–80. doi:10.1016/j.neuroimage.2006.01.021

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26, 297–302.

Doring, T.M., Kubo, T.T.A., Cruz, L.C.H., Juruena, M.F., Fainberg, J., Domingues, R.C., Gasparetto, E.L., 2011. Evaluation of hippocampal volume based on MR imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. J. Magn. Reson. Imaging 33, 565–572. doi:10.1002/jmri.22473

Dundas, E.M., Plaut, D.C., Behrmann, M., 2013. The joint development of hemispheric lateralization for words and faces. J. Exp. Psychol. Gen. 142, 348–358. doi:10.1037/a0029503

Glosser, G., Saykin, A.J., Deutsch, G.K., O'Connor, M.J., 1995. Neural organization of material-specific memory functions in temporal lobe epilepsy patients as assessed by the intracarotid amobarbital test. Neuropsychology 9, 449–456. doi:10.1037/0894-4105.9.4.449

Gogtay, N., Nugent, T.F., Herman, D.H., Ordonez, A., Greenstein, D.K., Hayashi, K.M., Clasen, L.S., Toga, A.W., Giedd, J.N., Rapoport, J.L., Thompson, P.M., 2006. Dynamic mapping of normal human hippocampal development. Hippocampus 16, 664–672. doi:10.1002/hipo

Golby, A.J., Poldrack, R.A., Brewer, J.B., Spencer, D., Desmond, J.E., Aron, A.P., Gabrieli, J.D.E., 2001. Material-specific lateralization in the medial temporal lobe and prefrontal cortex during memory encoding. Brain 124, 1841–1854. doi:10.1093/brain/124.9.1841

Gómez, R.L., Edgin, J.O., 2015. The extended trajectory of hippocampal development: Implications for early memory development and disorder. Dev. Cogn. Neurosci. 18, 57–69. doi:10.1016/j.dcn.2015.08.009

Gorgolewski, K.J., Poldrack, R. a, 2016. A practical guide for improving transparency and reproducibility in neuroimaging research. PLoS Biol. 1–13. doi:10.1371/journal.pbio.1002506

Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A., 2008. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. Neuroimage 40, 672–684. doi:10.1016/j.neuroimage.2007.11.034

Guo, T., Winterburn, J.L., Pipitone, J., Duerden, E.G., Park, M.T.M., Chau, V., Poskitt, K.J., Grunau, R.E., Synnes, A., Miller, S.P., Mallar Chakravarty, M., 2015. Automatic segmentation of the hippocampus for preterm neonates from early-in-life to term-equivalent age. NeuroImage Clin. 9, 176–193. doi:10.1016/j.nicl.2015.07.019

Kang, H.C., Burgund, E.D., Lugar, H.M., Petersen, S.E., Schlaggar, B.L., 2003. Comparison of functional activation foci in children and adults using a common stereotactic space. Neuroimage 19, 16–28. doi:10.1016/S1053-8119(03)00038-7

Kelley, W.M., Miezin, F.M., McDermott, K.B., Buckner, R.L., Raichle, M.E., Cohen, N.J., Ollinger, J.M., Akbudak, E., Conturo, T.E., Snyder, A.Z., Petersen, S.E., 1998. Hemispheric specialization in human dorsal frontal cortex and medial temporal lobe for verbal and nonverbal memory encoding. Neuron 20, 927–936. doi:10.1016/S0896-6273(00)80474-2

Krogsrud, S.K., Tamnes, C.K., Fjell, A.M., Amlien, I., Grydeland, H., Sulutvedt, U., Due-Tønnessen, P., Bjørnerud, A., Sølsnes, A.E., Håberg, A.K., Skrane, J., Walhovd, K.B., 2014. Development of hippocampal subfield volumes from 4 to 22 years. Hum. Brain Mapp. 5657, 5646–5657. doi:10.1002/hbm.22576

Lavenex, P., Banta Lavenex, P., 2013. Building hippocampal circuits to learn and remember: Insights into the development of human memory. Behav. Brain Res. doi:10.1016/j.bbr.2013.02.007

Lee, J.K., Ekstrom, A.D., Ghetti, S., 2014. Volume of hippocampal subfields and episodic memory in childhood and adolescence. Neuroimage 94, 162–171.

Lerma-Usabiaga, G., Iglesias, J.E., Insausti, R., Greve, D.N., Paz-Alonso, P.M., 2016. Automated segmentation of the human hippocampus along its longitudinal axis. Hum. Brain Mapp. 00. doi:10.1002/hbm.23245

Lin, M., Fwu, P.T., Buss, C., Davis, E.P., Head, K., Muftuler, L.T., Sandman, C.A., Su, M.Y., 2013. Developmental changes in hippocampal shape among preadolescent children. Int. J. Dev. Neurosci. 31, 473–481. doi:10.1016/j.ijdevneu.2013.06.001

Mack, M.L., Preston, A.R., 2016. Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. Neuroimage 127, 144–157. doi:10.1016/j.neuroimage.2015.12.015

Manns, J., Eichenbaum, H., 2006. Evolution of declarative memory. Hippocampus 16, 795–808. doi:10.1002/hipo

Martin, A., Wiggs, C.L., Weisberg, J., 1997. Modulation of human medial temporal lobe activity by form, meaning, and experience. Hippocampus 7, 587–593. doi:10.1002/(SICI)1098-1063(1997)7:6<587::AID-HIPO1>3.0.CO;2-C

Moser, M.B., Moser, E.I., 1998. Functional differentiation in the hippocampus. Hippocampus 8, 608–19. doi:10.1002/(SICI)1098-1063(1998)8:6<608::AID-HIPO3>3.0.CO;2-7

Mueller, S.G., Stables, L., Du, A.T., Schuff, N., Truran, D., Cashdollar, N., Weiner, M.W.,

2007. Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. Neurobiol. Aging 28, 719–26. doi:10.1016/j.neurobiolaging.2006.03.007

Olsen, R.K., Palombo, D.J., Rabin, J.S., Levine, B., Ryan, J.D., Rosenbaum, R.S., 2013. Volumetric analysis of medial temporal lobe subregions in developmental amnesia using high-resolution magnetic resonance imaging. Hippocampus 23, 855–860. doi:10.1002/hipo.22153

Pardoe, H.R., Pell, G.S., Abbott, D.F., Jackson, G.D., 2009. Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? Epilepsia 50, 2586–92. doi:10.1111/j.1528-1167.2009.02243.x

Pipitone, J., Park, M.T.M., Winterburn, J., Lett, T.A., Lerch, J.P., Pruessner, J.C., Lepage, M., Voineskos, A.N., Chakravarty, M.M., 2014. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. Neuroimage 101, 494–512. doi:10.1016/j.neuroimage.2014.04.054

Poppenk, J., Evensmoen, H.R., Moscovitch, M., Nadel, L., 2013. Long-axis specialization of the human hippocampus. Trends Cogn. Sci. 17, 230–40. doi:10.1016/j.tics.2013.03.005

Rodionov, R., Chupin, M., Williams, E., Hammers, A., Kesavadas, C., Lemieux, L., 2009. Evaluation of atlas-based segmentation of hippocampi in healthy humans. Magn. Reson. Imaging 27, 1104–1109. doi:10.1016/j.mri.2009.01.008

Schlichting, M.L., Guarino, K.F., Schapiro, A.C., Turk-Browne, N.B., Preston, A.R. Hippocampal structure predicts statistical learning and associative inference abilities during development. Manuscript under review.

Schlichting, M.L., Zeithamova, D., Preston, A.R., 2014. CA1 subfield contributions to memory integration and inference. Hippocampus 24, 1248–1260. doi:10.1002/hipo.22310

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

Schumann, C.M., Hamstra, J., Goodlin-Jones, B.L., Lotspeich, L.J., Kwon, H., Buonocore, M.H., Lammers, C.R., Reiss, A.L., Amaral, D.G., 2004. The Amygdala Is Enlarged in Children But Not Adolescents with Autism; the Hippocampus Is Enlarged at All Ages-Schumann et al. J. Neurosci. 24, 6392–6401. doi:10.1523/JNEUROSCI.1297-04.2004

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208–19.

doi:10.1016/j.neuroimage.2004.07.051

Strange, B.A., Witter, M.P., Lein, E.S., Moser, E.I., 2014. Functional organization of the hippocampal longitudinal axis. Nat. Rev. Neurosci. 15, 655–669. doi:10.1038/nrn3785

Tamnes, C.K., Walhovd, K.B., Engvig, A., Grydeland, H., Krogsrud, S.K., Østby, Y., Holland, D., Dale, A.M., Fjell, A.M., 2014. Regional Hippocampal Volumes and Development Predict Learning and Memory. Dev. Neurosci. 36, 161–174. doi:10.1159/000362445

Teicher, M.H., Andersen, S.L., Polcari, A., Anderson, C.M., Navalta, C.P., Kim, D.M., 2003. The neurobiological consequences of early stress and childhood maltreatment. Neurosci. Biobehav. Rev. 27, 33–44. doi:10.1016/S0149-7634(03)00007-1

Wechsler, D., 1999. Wechsler Abbreviated Scale of Intelligence. Psychological Corporation.

West, M.J., Gundersen, H.J., 1990. Unbiased stereological estimation of the number of neurons in the human hippocampus. J. Comp. Neurol. 296, 1–22. doi:10.1002/cne.902960102

Wisse, L.E.M., Biessels, G.J., Geerlings, M.I., 2014. A critical appraisal of the hippocampal subfield segmenation package in FreeSurfer. Front. Aging Neurosci. 6, 261. doi:10.1503/jpn.130070

Wisse, L.E.M., Gerritsen, L., Zwanenburg, J.J.M., Kuijf, H.J., Luijten, P.R., Biessels, G.J., Geerlings, M.I., 2012. Subfields of the hippocampal formation at 7T MRI: In vivo volumetric assessment. Neuroimage 61, 1043–1049. doi:10.1016/j.neuroimage.2012.03.023

Yushkevich, P. a., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S., 2010. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. Neuroimage 53, 1208–1224. doi:10.1016/j.neuroimage.2010.06.040

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., La Joie, R., Kerchner, G.A., LaRocque, K.F., Libby, L.A., Malykhin, N., Mueller, S.G., Olsen, R.K., Palombo, D.J., Parekh, M.B., Pluta, J.B., Preston, A.R., Pruessner, J.C., Ranganath, C., Raz, N., Schlichting, M.L., Schoemaker, D., Singh, S., Stark, C.E.L., Suthana, N., Tompary, A., Turowski, M.M., Van Leemput, K., Wagner, A.D., Wang, L., Winterburn, J.L., Wisse, L.E.M., Yassa, M.A., Zeineh, M.M., 2015a. 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. doi:10.1016/j.neuroimage.2015.01.004

Yushkevich, P.A., Avants, B.B., Pluta, J., Minkoff, D., Pickup, S., Liu, W., Gee, J.C.,

Grossman, M., Detre, A., 2008. A Computational Atlas of the Human Hippocampus from Postmortem Magnetic Resonance Imaging at 9.4 Tesla 56–57.

Yushkevich, P.A., Pluta, J.B., Wang, H., Xie, L., Ding, S.L., Gertje, E.C., Mancuso, L., Kliot, D., Das, S.R., Wolk, D.A., 2015b. Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. Hum. Brain Mapp. 36, 258–287. doi:10.1002/hbm.22627

Zeithamova, D., Preston, A.R., 2010. Flexible memories: Differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. J. Neurosci. 30, 14676 –14684. doi:10.1523/JNEUROSCI.3250-10.2010

TABLES

ROI	All DSC [95%CI]	Child DSC [95%CI]	Adolescent DSC [95%CI]	Adult DSC [95%CI]
ANTs				
HPC	0.84 [0.82 0.85]	0.83 [0.81 0.85]	0.83 [0.80 0.84]	0.85 [0.83 0.86]
CA_1	0.69 [0.68 0.70]	0.69 [0.67 0.71]	0.68 [0.66 0.70]	0.69 [0.68 0.71]
SUB	0.65 [0.64 0.66]	0.65 [0.63 0.67]	0.64 [0.62 0.65]	0.65 [0.62 0.68]
$DG/CA_{2/3}$	0.77 [0.76 0.78]	0.77 [0.75 0.79]	0.76 [0.74 0.79]	0.78 [0.78 0.79]
$CA_{2,3}$	0.43 [0.41 0.44]	0.42 [0.39 0.44]	0.42 [0.39 0.45]	0.44 [0.41 0.47]
DG	0.73 [0.72 0.74]	0.73 [0.71 0.74]	0.72 [0.70 0.74]	0.73 [0.72 0.75]
$ANTs_{ROI}$				
HPC	0.90 [0.89 0.90]	0.90 [0.89 0.90]*	0.90 [0.89 0.90]*	0.90 [0.90 0.91]
CA_1	0.73 [0.72 0.74]	0.74 [0.73 0.74]	0.72 [0.71 0.73]	0.73 [0.72 0.74]
SUB	0.73 [0.72 0.73]	0.73 [0.72 0.74]	0.72 [0.70 0.73]	0.73 [0.72 0.74]
$DG/CA_{2/3}$	0.81 [0.80 0.81]	0.81 [0.80 0.82]	0.80 [0.79 0.81]	0.81 [0.81 0.82]
$CA_{2,3}$	0.48 [0.47 0.49]	0.47 [0.45 0.49]	0.47 [0.46 0.49]	0.50 [0.48 0.51]
DG	0.75 [0.74 0.75]	0.75 [0.74 0.75]	0.74 [0.74 0.75]	0.75 [0.74 0.76]
$ASHS_S$				
HPC	0.93 [0.92 0.94]	0.92 [0.91 0.94]	0.93 [0.92 0.95]	0.94 [0.93 0.95]
CA_1	0.83 [0.82 0.85]	0.83 [0.79 0.86]	0.84 [0.81 0.87]	0.84 [0.81 0.86]
SUB	0.81 [0.80 0.83]	0.81 [0.77 0.85]	0.81 [0.78 0.85]	0.83 [0.80 0.85]
$DG/CA_{2/3}$	--	--	--	--
$CA_{2,3}$	0.69 [0.66 0.71]	0.66 [0.61 0.72]	0.70 [0.65 0.74]	0.70 [0.66 0.74]
DG	0.86 [0.85 0.87]	0.86 [0.84 0.88]	0.86 [0.84 0.89]	0.86 [0.84 0.88]
$ASHS_C$				
HPC	0.93 [0.92 0.94]	0.93 [0.91 0.94]	0.93 [0.92 0.94]	0.94 [0.93 0.95]
CA_1	0.83 [0.82 0.85]	0.83 [0.80 0.86]	0.84 [0.81 0.87]	0.84 [0.81 0.86]
SUB	0.82 [0.80 0.84]	0.81 [0.77 0.84]	0.81 [0.78 0.85]	0.83 [0.80 0.86]
$DG/CA_{2/3}$	0.89 [0.88 0.90]	0.88 [0.86 0.91]	0.89 [0.87 0.91]	0.89 [0.87 0.91]
$CA_{2,3}$	--	--	--	--
DG	--	--	--	--

Table 1. Spatial overlap of each method with Manual ROIs. Mean DSC and lower and upper bounds of 95% confidence intervals across all participants (left), as well as for child, adolescent, and adult groups separately. Asterisks indicate significance level of nonparametric t-tests comparing child and adolescent groups, respectively, with adults. * $p < 0.05$, uncorrected. See also **Figure 2**.

ROI	All ICC [95%CI]	Child ICC [95%CI]	Adolescent ICC [95%CI]	Adult ICC [95%CI]
ANTs				
HPC	0.76 [0.63 0.84]	0.76 [0.47 0.86]	0.76 [0.40 0.89]	0.77 [0.51 0.88]
CA_1	0.78 [0.68 0.85]	0.81 [0.57 0.90]	0.75 [0.37 0.88]	0.80 [0.64 0.88]
SUB	0.76 [0.62 0.86]	0.69 [0.27 0.86]	0.79 [0.50 0.91]	0.81 [0.50 0.95]
$DG/CA_{2/3}$	0.79 [0.68 0.86]	0.81 [0.55 0.91]	0.71 [0.36 0.88]	0.74 [0.50 0.86]
$CA_{2,3}$	0.56 [0.39 0.67]	0.53 [0.21 0.66]	0.62 [0.34 0.81]	0.51 [0.07 0.75]
DG	0.75 [0.61 0.84]	0.81 [0.43 0.96]	0.59 [0.02 0.84]	0.71 [0.49 0.83]
$ANTs_{ROI}$				
HPC	0.91 [0.87 0.94]	0.95 [0.89 0.97]	0.85 [0.73 0.93]	0.91 [0.84 0.95]
CA_1	0.90 [0.85 0.93]	0.92 [0.79 0.96]	0.88 [0.75 0.94]	0.90 [0.86 0.93]
SUB	0.85 [0.76 0.91]	0.89 [0.71 0.95]	0.84 [0.70 0.91]	0.78 [0.41 0.91]
$DG/CA_{2/3}$	0.77 [0.65 0.85]	0.88 [0.75 0.95]*	0.64 [0.27 0.83]	0.58 [0.17 0.78]
$CA_{2,3}$	0.54 [0.35 0.67]	0.56 [0.14 0.71]	0.58 [0.25 0.80]	0.38 [-0.21 0.70]
DG	0.70 [0.55 0.80]	0.88 [0.68 0.94]*	0.46 [-0.01 0.71]	0.55 [0.12 0.75]
$ASHS_S$				
HPC	0.91 [0.83 0.96]	0.89 [0.71 0.97]	0.90 [0.77 0.96]	0.97 [0.92 0.99]
CA_1	0.90 [0.82 0.94]	0.86 [0.69 0.96]	0.91 [0.84 0.95]	0.93 [0.86 0.97]
SUB	0.87 [0.77 0.92]	0.88 [0.72 0.95]	0.82 [0.56 0.93]	0.89 [0.70 0.96]
$DG/CA_{2/3}$	--	--	--	--
$CA_{2,3}$	0.63 [0.47 0.76]	0.62 [0.39 0.78]	0.58 [0.24 0.76]	0.64 [0.23 0.81]
DG	0.92 [0.86 0.96]	0.95 [0.83 0.98]	0.84 [0.63 0.95]	0.94 [0.88 0.96]
$ASHS_C$				
HPC	0.91 [0.83 0.96]	0.89 [0.70 0.97]	0.90 [0.76 0.96]	0.96 [0.91 0.98]
CA_1	0.90 [0.84 0.94]	0.87 [0.72 0.96]	0.91 [0.84 0.95]	0.92 [0.85 0.96]
SUB	0.86 [0.77 0.92]	0.88 [0.71 0.95]	0.81 [0.49 0.93]	0.89 [0.73 0.96]
$DG/CA_{2/3}$	0.90 [0.82 0.95]	0.90 [0.76 0.95]	0.87 [0.69 0.96]	0.93 [0.85 0.97]
$CA_{2,3}$	--	--	--	--
DG	--	--	--	--

Table 2. Volume correspondence for each automated method with Manual ROIs. Mean ICC value and lower and upper bounds of 95% confidence intervals across all participants (left), as well as for child, adolescent, and adult groups separately. Asterisks indicate significance level of nonparametric t-tests comparing child and adolescent groups, respectively, with adults. * $p < 0.05$, uncorrected. See also **Figure 4**.

ROI	Volume F statistic (p value)	Age F statistic (p value)	Volume x Age F statistic (p value)
ANTs			
HPC	18.21 (0.002)**	4.12 (0.06)	0.37 (0.51)
CA_1	43.54 (<0.001)**	8.11 (0.003)**	0.60 (0.46)
SUB	0.96 (0.50)	2.04 (0.21)	0.59 (0.54)
$DG/CA_{2/3}$	38.35 (<0.001)**	3.27 (0.08)	0.62 (0.47)
$CA_{2,3}$	116.20 (<0.001)**	3.75 (0.06)	1.26 (0.27)
DG	20.17 (<0.001)**	1.89 (0.19)	1.29 (0.26)
$ANTs_{ROI}$			
HPC	12.31 (0.03)*	4.71 (0.02)*	1.79 (0.27)
CA_1	38.31 (<0.001)**	5.48 (0.02)*	1.67 (0.19)
SUB	3.88 (0.17)	1.71 (0.22)	0.49 (0.46)
$DG/CA_{2/3}$	12.72 (0.03)*	1.14 (0.29)	1.69 (0.26)
$CA_{2,3}$	95.92 (<0.001)**	3.17 (0.09)	1.21 (0.29)
DG	3.79 (0.19)	0.66 (0.36)	1.76 (0.24)
$ASHS_S$			
HPC	18.76 (0.005)**	4.52 (0.01)*	1.22 (0.33)
CA_1	34.06 (<0.001)**	2.65 (0.10)	0.82 (0.36)
SUB	2.17 (0.30)	1.73 (0.20)	3.94 (0.03)*
$DG/CA_{2/3}$	--	--	--
$CA_{2,3}$	77.88 (<0.001)**	9.81 (0.002)**	0.28 (0.67)
DG	5.01 (0.11)	0.98 (0.31)	2.20 (0.20)
$ASHS_C$			
HPC	17.58 (0.004)**	4.64 (0.007)**	1.12 (0.35)
CA_1	43.07 (<0.001)**	3.41 (0.05)	0.63 (0.40)
SUB	1.94 (0.33)	1.63 (0.20)	3.36 (0.05)
$DG/CA_{2/3}$	17.60 (0.02)*	3.02 (0.10)	1.81 (0.22)
$CA_{2,3}$	--	--	--
DG	--	--	--

Table 3. Bias statistics for each automated method, with Manual ROIs serving as the baseline. F statistics and corresponding p values indicate the reliability of the main effects of volume, age, and volume x age interactions. While all methods showed evidence of significant bias across the group for some regions, there is little evidence that these biases are impacted by developmental stage. * $p < 0.05$ and ** $p < 0.01$, uncorrected. These data are depicted in **Figures 5-8**.

ROI	All	Child	Adolescent	Adult
Manual				
HPC	0.93 [0.90 0.96]	0.95 [0.89 0.98]	0.89 [0.79 0.96]	0.95 [0.90 0.98]
CA_1	0.91 [0.87 0.94]	0.88 [0.76 0.95]	0.91 [0.86 0.96]	0.95 [0.89 0.98]
SUB	0.65 [0.50 0.77]	0.77 [0.50 0.91]	0.55 [0.25 0.80]	0.53 [0.20 0.76]
$DG/CA_{2,3}$	0.79 [0.67 0.88]	0.88 [0.70 0.96]	0.65 [0.33 0.85]	0.74 [0.49 0.88]
$CA_{2,3}$	0.71 [0.56 0.81]	0.83 [0.62 0.93]	0.49 [0.16 0.73]	0.61 [0.42 0.79]
DG	0.78 [0.68 0.86]	0.85 [0.62 0.94]	0.77 [0.43 0.92]	0.72 [0.47 0.89]
ANTs				
HPC	0.88 [0.81 0.93]	0.86 [0.72 0.95]	0.85 [0.75 0.92]	0.89 [0.66 0.96]
CA_1	0.81 [0.69 0.89]	0.83 [0.68 0.94]	0.79 [0.53 0.92]	0.70 [0.37 0.91]
SUB	0.62 [0.48 0.75]	0.76 [0.57 0.87]	0.48 [0.18 0.73]	0.51 [0.11 0.79]
$DG/CA_{2,3}$	0.80 [0.69 0.88]	0.92 [0.80 0.97]*	0.78 [0.64 0.89]	0.54 [0.20 0.81]
$CA_{2,3}$	0.53 [0.29 0.71]	0.82 [0.59 0.93]**	0.38 [-0.06 0.69]	0.24 [-0.31 0.62]
DG	0.84 [0.76 0.90]	0.91 [0.77 0.96]	0.84 [0.70 0.94]	0.68 [0.44 0.85]
$ANTs_{ROI}$				
HPC	0.92 [0.89 0.95]	0.95 [0.88 0.98]	0.89 [0.79 0.95]	0.89 [0.78 0.95]
CA_1	0.85 [0.78 0.91]	0.88 [0.73 0.96]	0.81 [0.67 0.91]	0.85 [0.72 0.93]
SUB	0.74 [0.63 0.82]	0.78 [0.49 0.92]	0.65 [0.34 0.82]	0.78 [0.61 0.90]
$DG/CA_{2,3}$	0.81 [0.70 0.89]	0.91 [0.79 0.96]*	0.83 [0.58 0.96]	0.62 [0.39 0.83]
$CA_{2,3}$	0.62 [0.41 0.77]	0.78 [0.57 0.91]	0.68 [0.37 0.87]	0.29 [-0.28 0.72]
DG	0.84 [0.74 0.90]	0.90 [0.76 0.96]*	0.87 [0.63 0.96]	0.69 [0.45 0.86]
$ASHS_S$				
HPC	0.94 [0.91 0.96]	0.93 [0.84 0.97]	0.93 [0.84 0.97]	0.94 [0.85 0.97]
CA_1	0.83 [0.75 0.89]	0.79 [0.53 0.92]	0.85 [0.74 0.94]	0.91 [0.78 0.97]
SUB	0.72 [0.57 0.82]	0.78 [0.49 0.90]	0.44 [0.05 0.72]*	0.86 [0.67 0.95]
$DG/CA_{2,3}$	--	--	--	--
$CA_{2,3}$	0.53 [0.32 0.71]	0.63 [0.29 0.87]*	0.42 [0.05 0.68]	0.10 [-0.26 0.50]
DG	0.86 [0.79 0.91]	0.92 [0.84 0.97]	0.88 [0.78 0.94]	0.75 [0.51 0.90]
$ASHS_C$				
HPC	0.93 [0.90 0.96]	0.93 [0.82 0.98]	0.93 [0.86 0.97]	0.92 [0.84 0.97]
CA_1	0.80 [0.70 0.86]	0.77 [0.47 0.92]	0.84 [0.73 0.93]	0.85 [0.63 0.97]
SUB	0.74 [0.61 0.84]	0.81 [0.56 0.92]	0.48 [0.12 0.75]*	0.85 [0.63 0.95]
$DG/CA_{2,3}$	0.83 [0.74 0.89]	0.90 [0.79 0.96]*	0.86 [0.75 0.94]*	0.63 [0.34 0.82]
$CA_{2,3}$	--	--	--	--
DG	--	--	--	--

Table 4. Within-method volume correspondence for all methods, indexed as Pearson correlations between left and right hemisphere volumes. Lower and upper bounds indicate 95% confidence intervals across all participants (left), as well as for child, adolescent, and adult groups separately. Asterisks indicate significance level of nonparametric t-test comparing child and adolescent groups, respectively, with adults. * $p < 0.05$ and ** $p < 0.01$, uncorrected. See also **Figure 9**.

FIGURES AND CAPTIONS

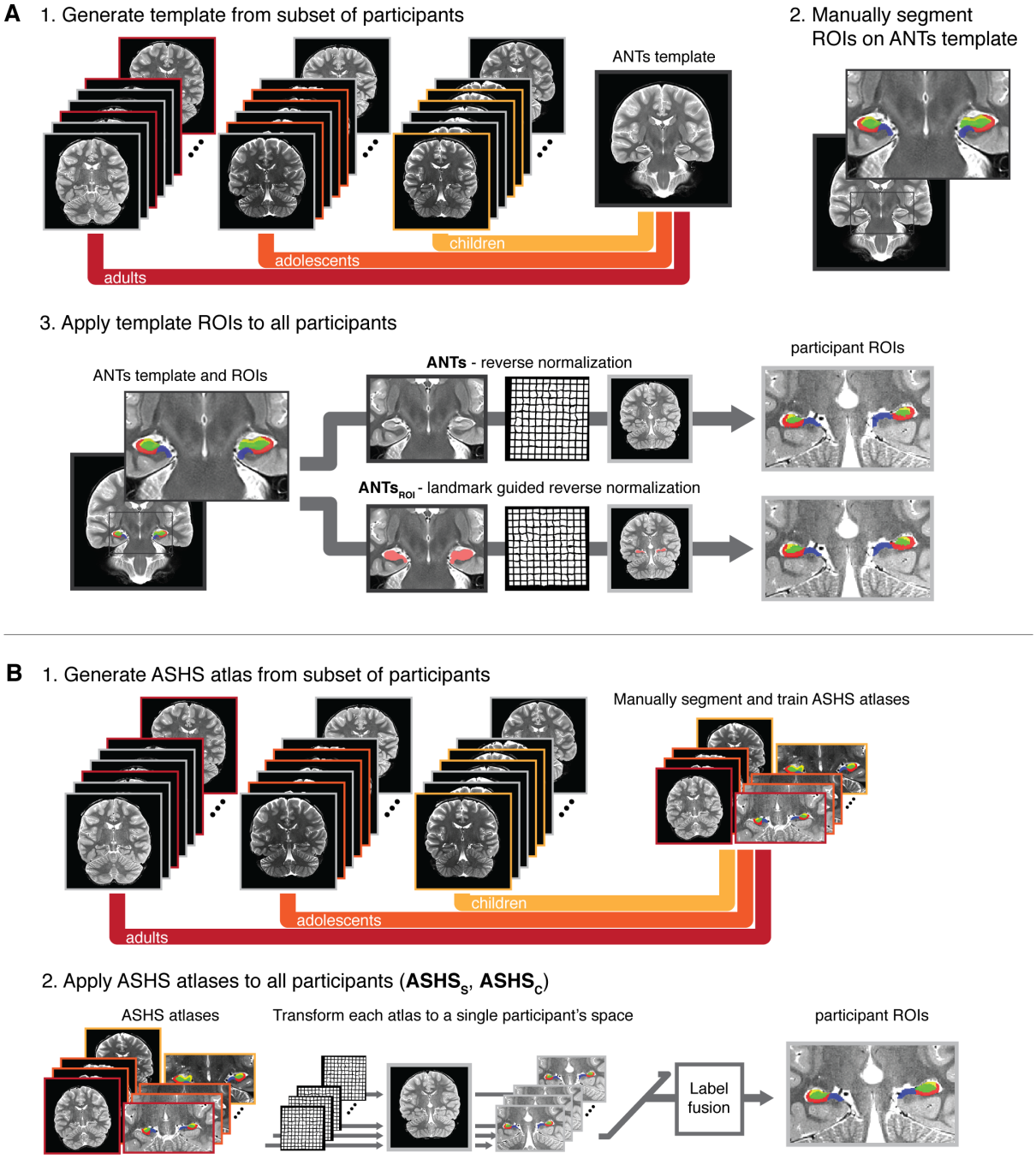

Figure 1. Schematic depiction of methods compared. **A)** ANTs segmentation methods. **1.** Subsets of participants were selected from the three age groups to generate a single ANTs template. **2.** Then, hippocampal subfield ROIs were manually delineated on the ANTs template. **3.** Finally, the ANTs template ROIs were reverse normalized to each participant's native space. For the **ANTs** method (top path), a nonlinear warp was estimated based uniformly on the entire ANTs template and participant anatomical volumes. For the **ANTs$_{ROI}$** method (bottom path), the nonlinear warp was estimated through landmark matching of the whole hippocampus. Both

ANTs-based methods were performed for all participants to create two sets of participant-specific subfield ROIs. **B)** ASHS segmentation methods. **1.** Subsets of participants from the three age groups were selected for training custom ASHS atlases. The ASHS training procedure was run using the manually delineated subfield segmentations for the atlas participants. **2.** Then, the ASHS atlases were applied to all participants. Briefly, the ROIs from each atlas participant were all nonlinearly warped to a single participant. Then, a label fusion procedure combined the transformed atlas segmentations to generate final, participant-specific ROIs. The whole procedure was performed twice: once with DG and $CA_{2,3}$ separated (**$ASHS_S$**), and once with them combined (**$ASHS_C$**).

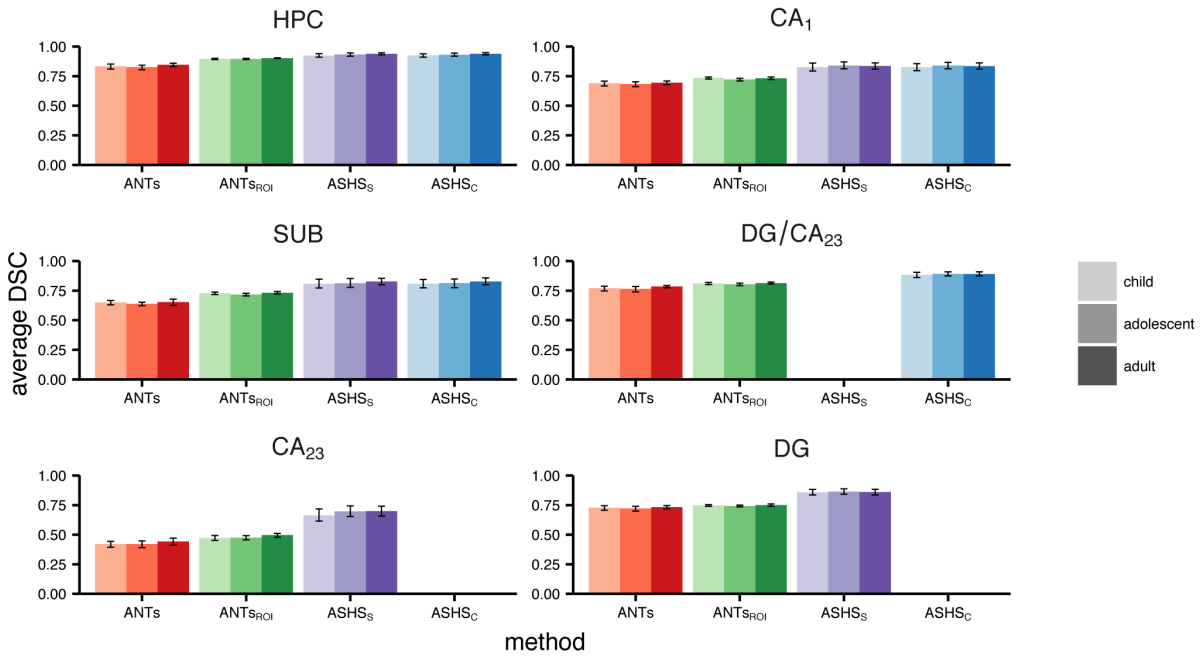

Figure 2. Spatial overlap of each method with Manual ROIs measured using DSC. Error bars represent 95% confidence intervals. Data correspond with **Table 1**.

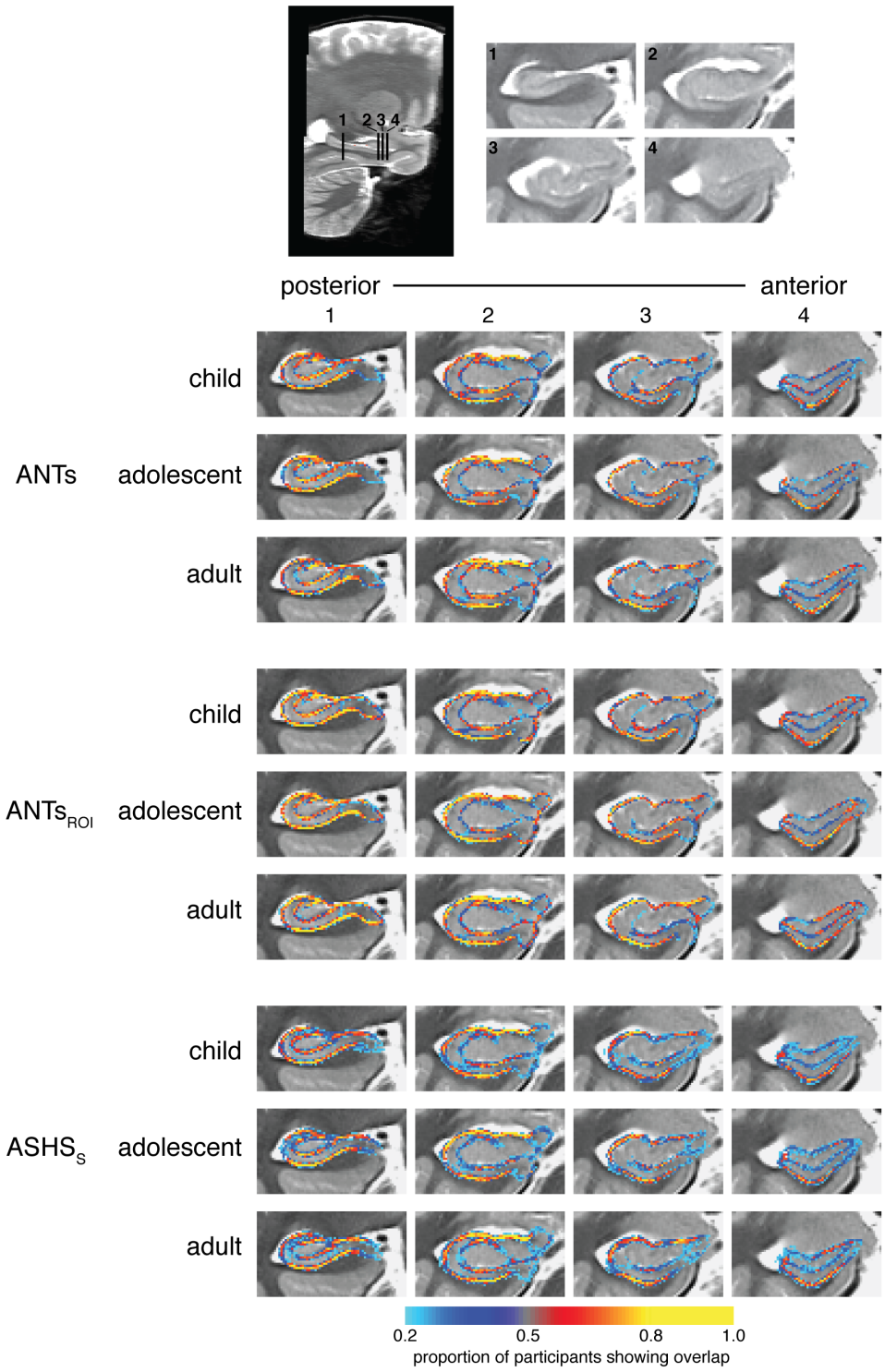

Figure 3. Voxelwise edge agreement displayed on a custom template separately for children, adolescents, and adults. Intensities represent the proportion of participants for which the method and Manual agreed that the voxel was a subfield boundary.

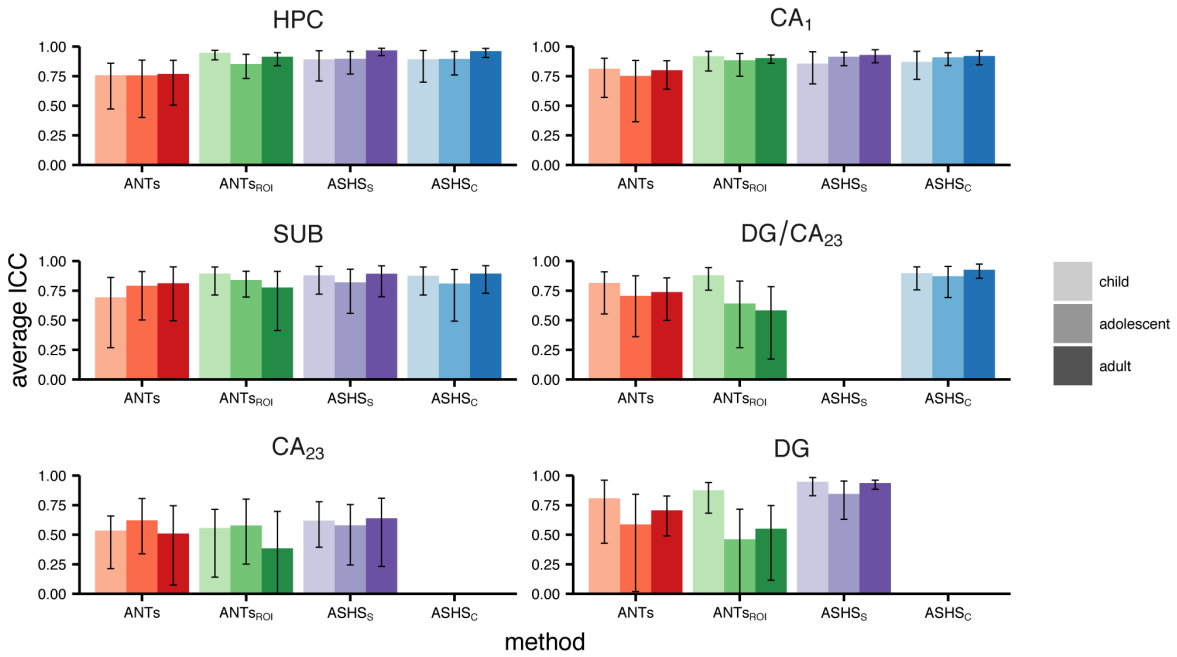

Figure 4. Volume correspondence of each method with Manual ROIs measured using ICC. Error bars represent 95% confidence intervals. Data correspond with **Table 2**.

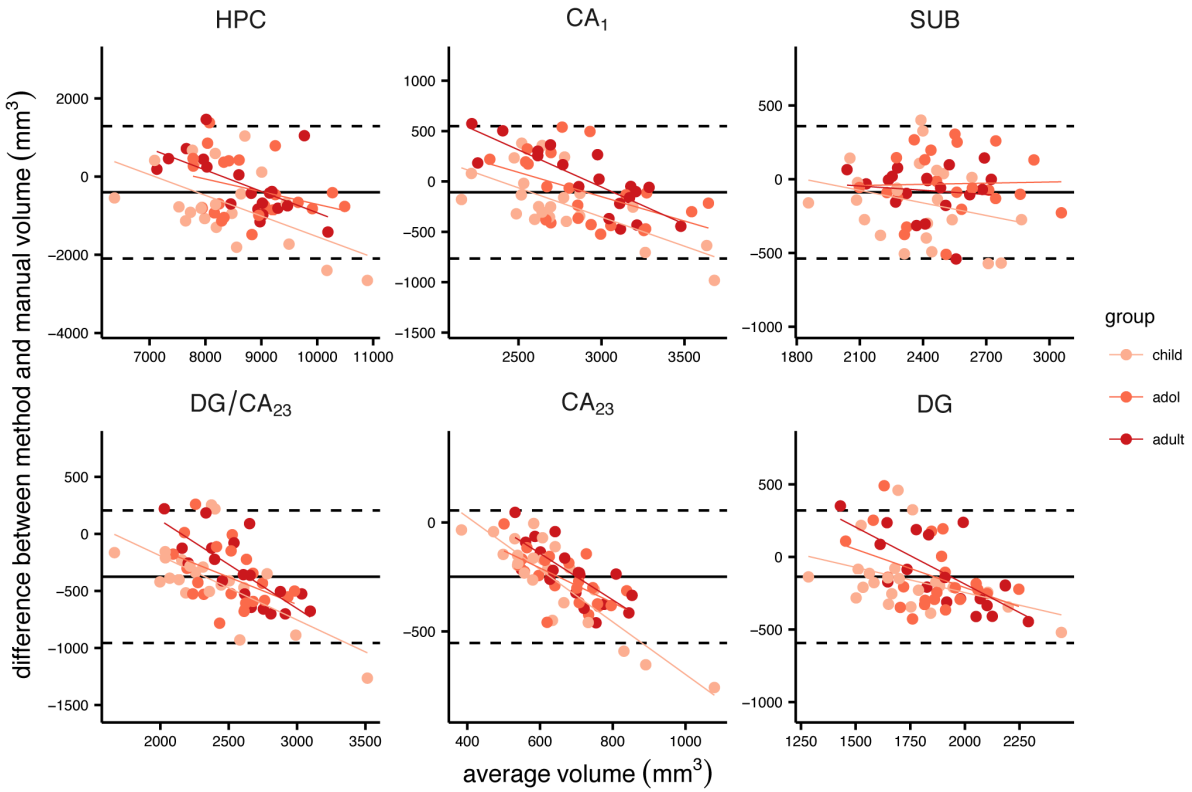

Figure 5. Bland-Altman plots comparing ANTs with Manual. X-axis represents the mean regional volume across the two methods; y-axis represents the difference (ANTs-Manual). Solid black line indicates the mean difference across all age groups; dashed lines are 2 standard deviations above and below the mean. Regression lines are displayed for each age group separately. ANCOVA statistics are provided in **Table 3**.

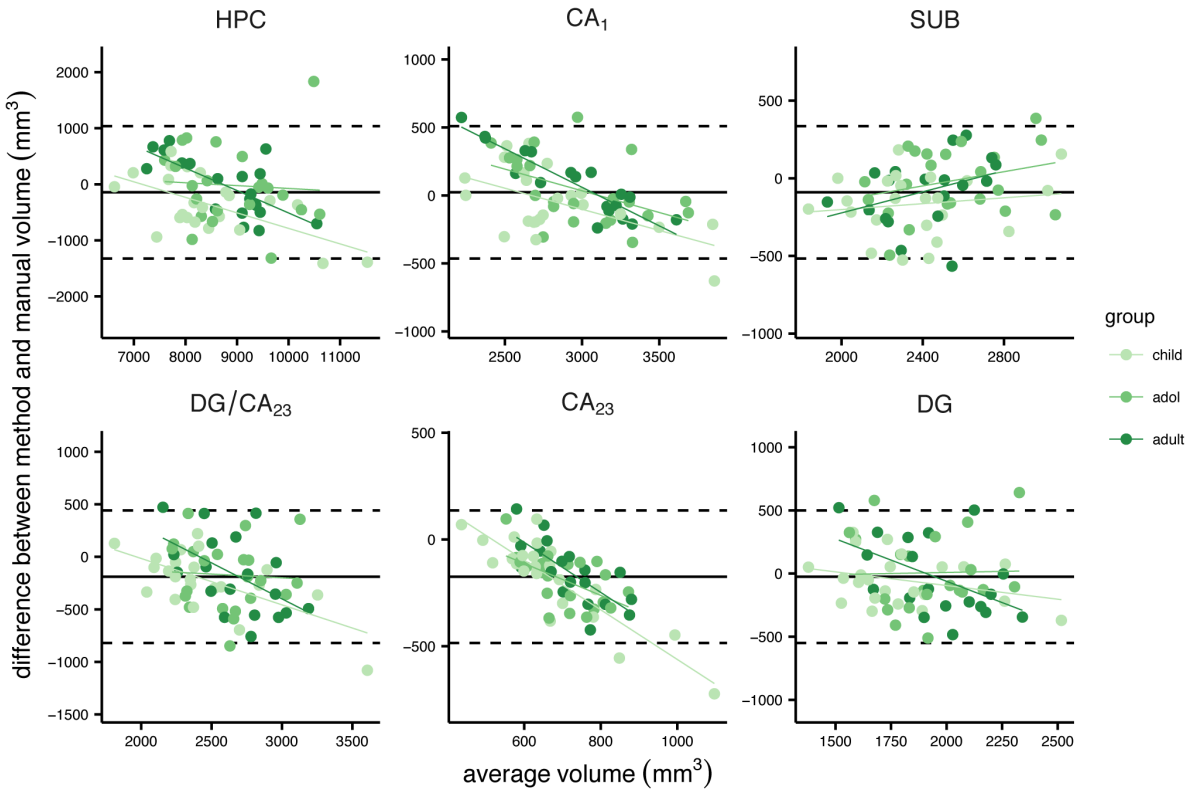

Figure 6. Bland-Altman plots comparing $ANTs_{ROI}$ with Manual. Data are displayed as in **Figure 5**. ANCOVA statistics are provided in **Table 3**.

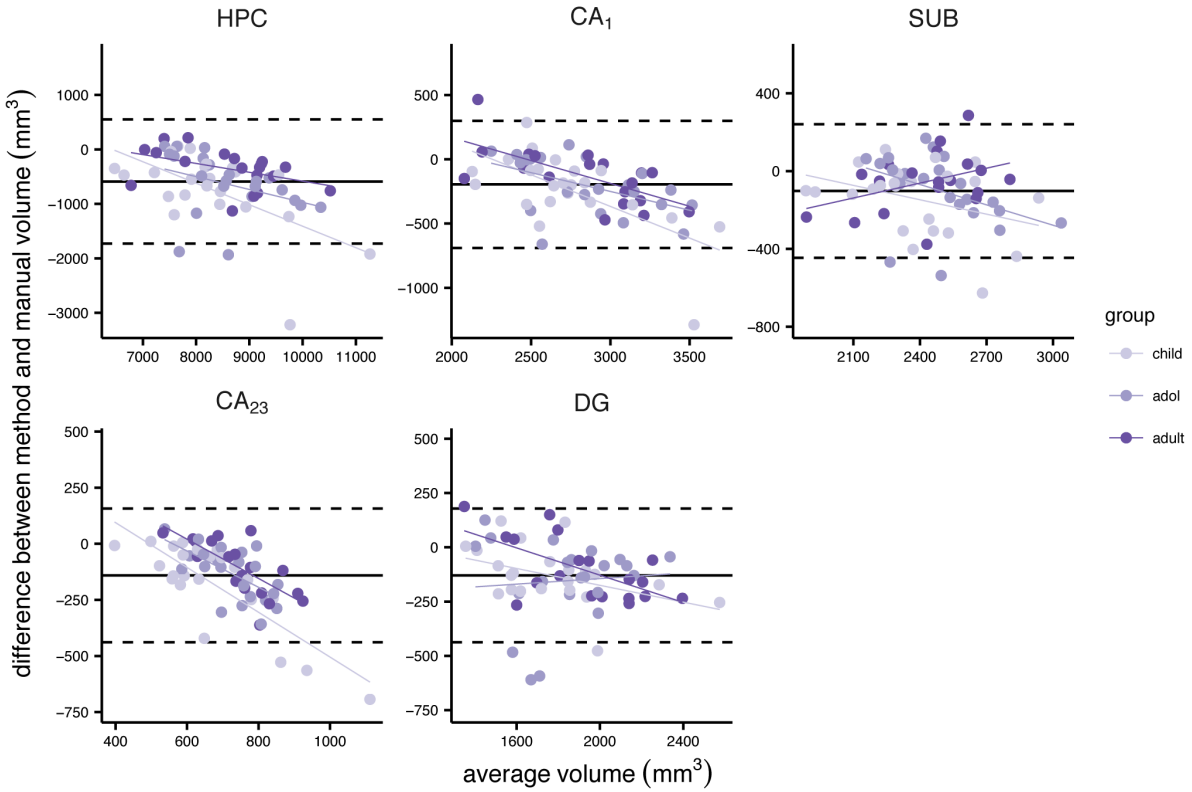

Figure 7. Bland-Altman plots comparing $ASHS_S$ with Manual. Data are displayed as in **Figure 5**. ANCOVA statistics are provided in **Table 3**.

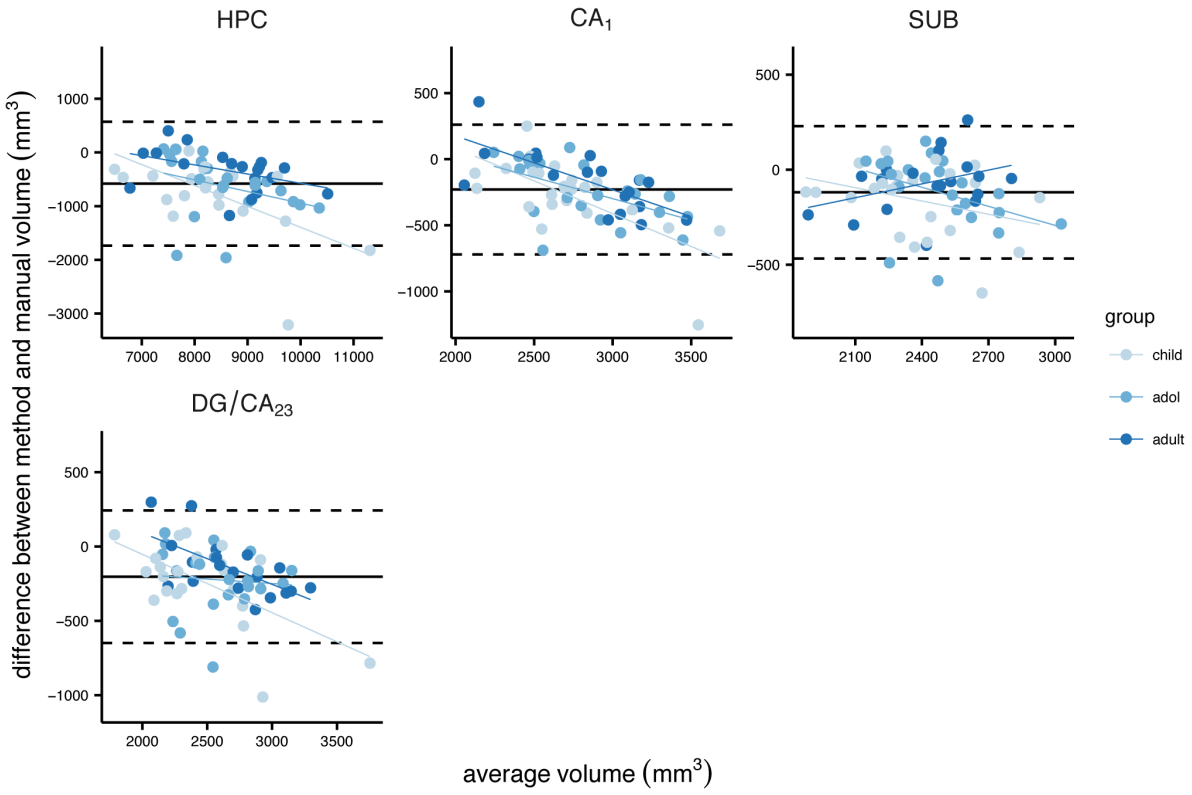

Figure 8. Bland-Altman plots comparing $ASHS_C$ with Manual. Data are displayed as in **Figure 5**. ANCOVA statistics are provided in **Table 3**.

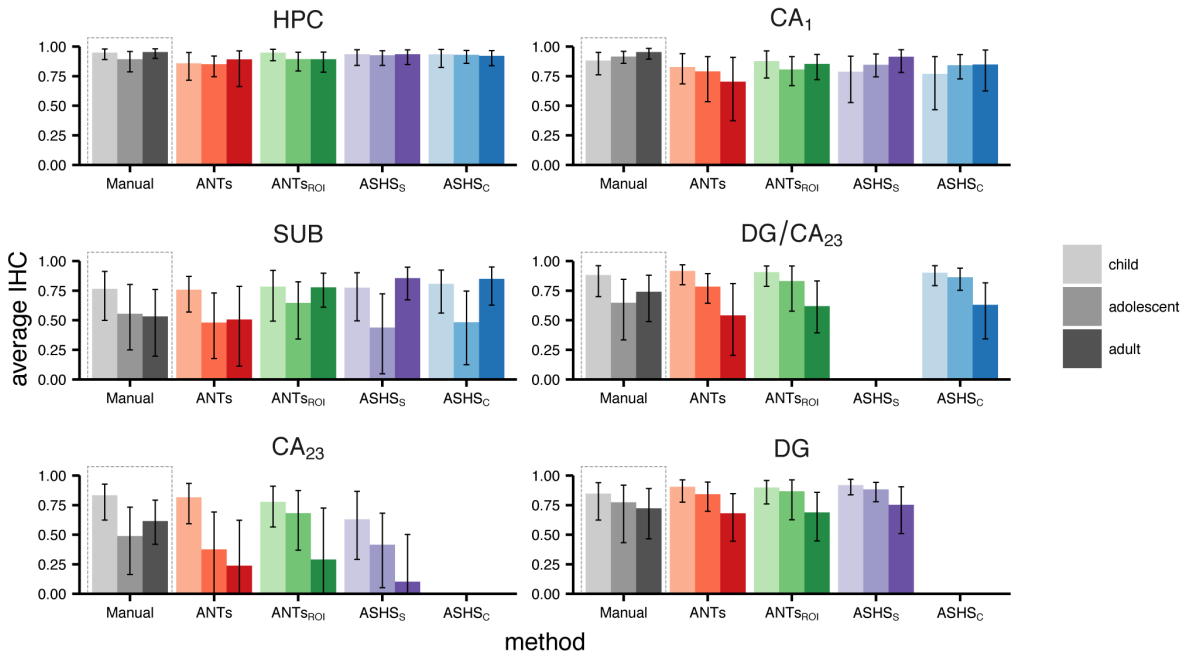

Figure 9. Within-method reliability assessed using IHC. Bar height represents the Pearson's r value between the left and right hemisphere volumes for each group separately; error bars denote 95% confidence intervals. Data correspond with **Table 4**.