Evaluation of Automatic and Manual Segmentations in Neonate Subcortical Structures

Kirsten Nicole Zaldarriaga Consing[a], Claudia Buss[b,c], Pathik D Wadhwa[c], Jeffrey T Young[a,d], Sonja Entringer[b,c], Mehdi Sarkeshi[f], Cheryl Dietrich[e], Audrey R Verde[a], John H. Gilmore[a], Jerod Rasmussen[c], Martin Andreas Styner[a,d]

[a] *Department of Psychiatry, University of North Carolina at Chapel Hill, NC, USA*
[b] *Department of Medical Psychology, Charité University Medicine Berlin, Germany*
[c] *Development, Health and Disease Research Program, University of California Irvine, Irvine, USA*
[d] *Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA*
[e] *Department of Epidemiology, University of Washington, Seattle Wa, USA*
[f] *Wolfram Research, Champaign IL, USA*

ARTICLE INFO

Article history:

Keywords:
Neuroimaging
Compulational Atlases
Magnetic Resonance Imaging
Automatic Segmentation

ABSTRACT

Segmentation of the hippocampus and amygdala is a crucial task in studies that aim at understanding the role of inter-individual variation in the size of these structures in neurodevelopmental and neurodegenerative disorders. While methods for automatic segmentation of these structures are available, manual segmentation is generally still considered the golden standard, especially in pediatric samples due to the relatively low quality of automatic hippocampal and amygdala segmentations, especially in areas of low contrast close to the neighboring regions of the amygdala and hippocampus. Here, we evaluate the differences between automatic and manually edited segmentations in neonate subjects for an automatic segmentation that employs a multi-modality, multi-atlas approach. The evaluation is based on coefficients of variance (COV), intra-class and inter-class correlations, volumetric overlap and surface distance metrics for both structures as well as local shape analysis for the hippocampus. We found a high degree of reliability within and across raters for the manual editing of the automatic segmentations. Surprisingly, the left hippocampus underwent a significantly higher amount of editing compared to the right. Furthermore, none of the structures show a significant asymmetric manual rater bias due to presentation mode. Via scan/rescan datasets, the automatic segmentations of the hippocampus and amygdala were observed to be appropriately reliable. In addition, we find that for both hemispheres, the amygdalae and hippocampi display a high degree of volumetric overlap (>98%) and low average surface distance (<1mm) between automatic and manually edited version. Shape analysis of the hippocampus shows mainly regions of reduction, mostly located close to the hippocampus-amygdala boundary. Our results overall indicate that manual editing yields consistently different segmentations from the automatic ones, though with relatively little benefit given the the high degree of agreement between the two.

1. Introduction

The hippocampus and amygdala are subcortical structures of relevance in many neuroimaging studies. The task of segmenting these structures is crucial in such studies, for example, in order to understand the association between alterations in anatomy of the hippocampus and amygdala and neurodevelopmental disorders (Hajek et al, 2009; Dager et al, 2007; Frodi et al, 2010; Kesler et al, 2004). Similarly, neuroimaging studies in neurodegenerative diseases such as Alzheimer's disease and other types of dementia have illuminated the role of the amygdala (Lehmann et al, 2010; Pinkhardt et al, 2006) in its neuropathology. In addition, the hippocampus has been shown to be crucial to the understanding of several other neuro-developmental and degenerative diseases such as schizophrenia, Alzheimer's, dementia, and epilepsy (Chupin et al, 2007; Collins & Pruessner, 2010; Dill et al, 2015; Lötjönen et al, 2015; Inglese et al, 2015)

Due to the importance of these structures, numerous methods have been proposed for their automatic segmentation (Chupin et al 2007; Hanson et al, 2012; Hu et al, 2011). However, manual segmentation though is still employed in most neuroimaging studies as current automatic methods result in segmentations that are often judged as not fully appropriate without further correction. The major difficulty to automatic segmentation is the lack of significant contrast in several regions in the hippocampus and amygdala, particularly where they both border each other. The problem is exacerbated in pediatric scans where the contrast between white and gray matter, as well as between hippocampus and amygdala is further reduced. Historically, surface model-based and single atlas methods have been employed most frequently, though they have also been found to need significant manual correction for improvement (Morey et al, 2009; Schoemaker et al, 2016). Recently, multi-atlas approaches have started to yield significantly improved automatic segmentations in

comparison to prior methods in. (Alijabar et al, 2009; Gholipour et al, 2012; Hanson et al, 2012; Lötjönen et al, 2010; Rohlfing et al, 2003; Wang et al, 2014). These multi-atlas based segmentation methods have provided significant inroads towards anatomically acceptable segmentation of these structures and are now increasingly employed in neuroimaging studies in adults (Lötjönen et al, 2010), though no application to the neonate setting has been shown yet.

In this paper, we focus on MR images from neonate subjects, where the lack of contrast and signal-to-noise ratio is pronounced. As a result, few neuroimaging studies of neonatal amygdala and hippocampus anatomy (Buss et al, 2012; Thompson et al, 2008) exist. The low contrast seen in this setting can be attributed to the low degree of myelination in white matter abutting the hippocampus. Manual segmentations of those neonatal structures are thus common, but highly time-consuming, taking hours to complete from scratch. Such manual segmentations also have relatively high intra-rater and inter-rater variability. Consequently, automatic methods with a high degree of segmentation quality would be desirable to alleviate these issues. This paper presents one such segmentation method in the neonate MRI setting using a multi-modality, multi-atlas approach that shows reliability, stability and appropriate accuracy in a study featuring 89 neonate subjects. It is worthwhile to note that the method presented here is publically available, both the tools and atlases.

2. Materials and Methods

2.1 Subjects

Infant neuroimaging was approved by the Institutional Review Board of the University of California at Irvine, and all parents provided informed, written consent. All of the 89 infants evaluated in this study were from healthy pregnancies with no major obstetric, birth or current health complications. Gestational age was determined by best obstetric estimate with a combination of last menstrual period and early uterine size, and was confirmed by obstetric ultrasonographic biometry before 15 weeks using standard clinical criteria (O'Brien et al., 1981). The mean gestational age at birth was 39.1 ± 1.6 (±SD) weeks and ranged from 34.4 to 41.9 weeks. The mean postnatal infant age at assessment was 25.5 ± 12.2 (±SD) days and ranged from 5 to 56 days.

2.1.1 MRI Acquisition

MRI scans were acquired during natural sleep using a 12-channel head receive coil on a 3T Siemens Tim Trio scanner. After feeding and soothing to the point of sleep, neonates were placed in a CIVCO beaded pillow (www.civco.com). The pillow covered the neonates' body and head, became rigid under vacuum, and provided a comforting swaddle, motion prevention and hearing protection in conjunction with foam earplugs. A pediatric specialist observed the neonates throughout the duration of scans, monitoring for heart rate and oxygen saturation via a pulse-oximeter attached to the foot. The entire protocol included T1-weighted, T2-weighted, diffusion tensor and functional imaging of the brain. The high-resolution anatomical scans consisted of a T1-weighted (MPRAGE, TR/TE/TI= 2400/ 3.16/ 1200ms, Flip Angle=8 degrees, Matrix= 256x256x160, Resolution=1x1x1mm, 6m 18s) and T2-weighted (TSE, TR/TE=3200/255ms, Matrix= 256x256x160, Resolution=1x1x1mm, 4m 18s) scan.

2.1.2 Multi-Atlas Population

We used a neonate multi-atlas population dataset consisting of 8 subjects with good image quality selected from the subject data described above. These subjects were manually segmented without prior starting segmentation by an expert rater at the UNC Neuro Image Research and Analysis Lab (CD) in a hippocampal long-axis aligned view. In order to allow for unbiased asymmetry analysis, all 8 subject MRIs and segmentation data were mirrored, resulting in an overall multi-atlas population of 16 atlases. This atlas has been made publically available (see resource section).

2.1.3 Subjects for Reliability Analysis

The segmentation reliability of the automatic multi-atlas segmentation was evaluated with 6 datasets consisting each of 2 scan sets acquired at the same scanning visit. The six subjects were selected from the subject population described above (2 subjects), as well as another 4 subjects from the UNC early brain development database (Gilmore et al, 2012). All scans were acquired on 3T Siemens Tim Trio scanners at UNC and UCI. For these subjects a second set of T1weighted and T2weighted scans were acquired in the same scanning session, as the first set was considered of borderline quality by the scanning technicians. In all these cases, subsequent quality control procedures by a trained image analysis expert (MAS) showed all scans to pass quality assessment for structural morphometric analysis. This small size neonate scan-rescan database captures the low signal-to-noise setting and the presence of motion very common in early postnatal scans. It is thus well suited to estimate the reliability of image processing procedures of neonate MRI data. As mentioned, at least half of these images for each scan session were assessed to be of borderline quality by the scanner personnel, and thus this scan-rescan evaluation is likely to overestimate the expected variability as compared to the average scan setting.

2.2 Multi-Modality Multi-Atlas Segmentation

Employing our automatic segmentation pipeline tool AutoSeg, we applied the following processing to the data: 1) inhomogeneity correction via N4 (Tustison, et al 2010), 2) rigid registration to a prior neonate atlas in ICBM space, 3) atlas moderated Expectation Maximization optimization based tissue classification for automatic skull stripping via ABC, 4) skull stripping using the tissue segmentation result from step (3) and finally 4) multi-atlas based structural segmentation (Wang, et. al 2014). Before running step (4), we manually correct all automatic brain masks employing the T1weighted image with a standard protocol removing only skull and exterior regions, while keeping other regions, such as the cerebellum and extra-axial CSF regions intact. During the multi-atlas segmentation step all atlases and subject MRI images are pair-wise co-registered, intensity and shape based similarity metrics are computed between scans and atlases, and a weighted majority voting based label fusion employing these metrics creates the final

segmentation (Wang, et. al 2014). In order to compute an automatic segmentation that is unbiased with respect to the cerebral hemisphere, we incorporated left-right mirrored versions of all atlas datasets, which in turn also duplicates the number of available atlases.

2.3 Manual Correction of Segmentations

The automatic segmentation results were edited via manual outlining on the T1 weighted image in all three orthogonal slice directions using the ITK-Snap (Yushkevich et al., 2006) segmentation tool. For the purpose of improved visual presentation during the manual correction, the data was first realigned such that the hippocampal long axis coincided with the anterior-posterior axis (see Figure 1). After manual editing, the inverse transform was applied to the edited segmentations to bring them back into neonate ICBM space.
As a way to test for a possible asymmetric rater bias, the multi-atlas segmentation is applied to the subject data presented in original orientation and to the same data but mirrored along the left-right axis

To assess intra and inter-rater variability, 5 subjects were randomly selected and triplicated for a total of 15 scans that were segmented by two raters (MS, KNZC). For the assessment of the automatic segmentation, a single rater segmented all 89 datasets (KNZC).

The appendix shows details of the manual editing process on the example of the right hippocampus. The rater begins in the axial plane through each slice of the brain image to search for voxels that are missing or incorrectly labeled from the segmentation in its 2D representation. If the rater identifies a voxel that needs to be removed or added, the rater assessed that voxel in the 3D rendering and along each of the three slices. After going through the axial plane, the rater continues the same method through the sagittal and coronal planes. As both hippocampus and amygdala are considered smoothly shaped, a smooth appearance was favored in this editing process, unless clear evidence to the contrary was present in the image.

2.4 Evaluation Metrics

2.4.1 Coefficient of Variance

The two trained human raters segmented each reliability dataset in its original and left-right mirrored presentation. The average coefficients of variance (COV) for all amygdala and hippocampal volume measurements were determined for each rater individually and then averaged over all raters. Coefficient of variance (standard deviation / mean value) explains the extent of variability between segmentations in relation to the average volume.

2.4.2 Intra-Class Correlation and Inter-Class Correlation

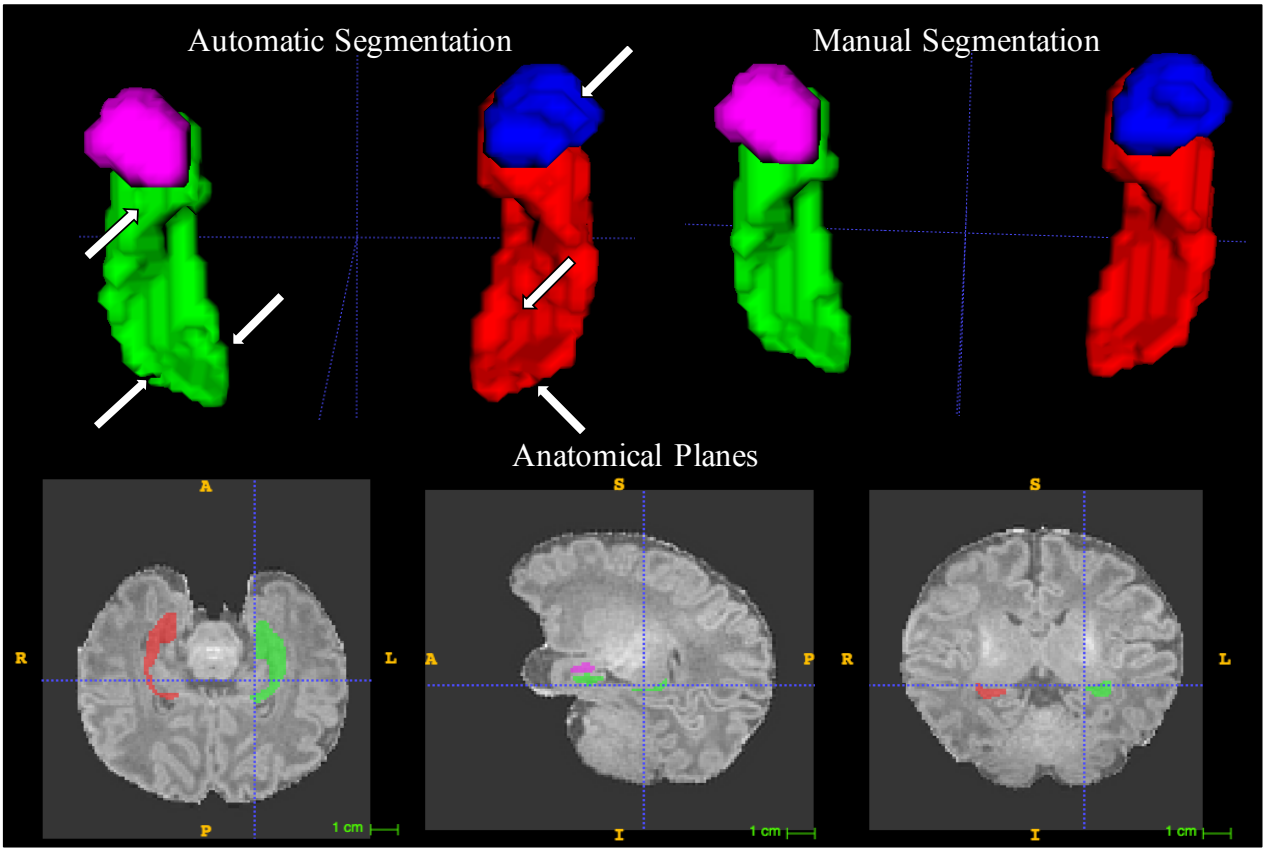


Fig 1. Automatic and manual segmentation of the hippocampus and amygdala performed on the hippocampal long-axis aligned MR images in ITK-SNAP (here shown with the T1-weighted image). Arrows indicate areas of manual editing. The hippocampi are is labeled red (right) and green (left), the amygdalae are labeled blue (right) and pink (left).

Intra-class as well as inter-class correlations were computed for both raters. Inter-class correlation is defined as the degree of agreement in volume of the edited segmentations between raters. Intra-class correlation is defined as the degree of agreement of multiple edited segmentations for a single rater.

2.4.3 Volumetric Overlap and Surface Distance (Mean/Max)

We assessed the performance of the multi-atlas segmentation method by evaluating how closely shaped the resulting segmentation is to the corresponding manually edited (reference) segmentation. The most commonly used metric in the field is the Dice similarity coefficient (DSC), which is computed between two segmentations as:

$$DSC = 2\,x\,\frac{V_{auto} \cap V_{ref}}{V_{auto} + V_{ref}}\,x\,100\%$$

where *V*auto and *V*ref are the volume of the automated segmentation result and the volume of the reference segmentation, respectively. In our case, *V*ref represents the volume of the manually edited segmentation. A DSC of 1 indicates complete volumetric overlap, and 0 indicates no overlap at all. We also employed the symmetric mean absolute distance (MAD) and the symmetric Hausdorff distance (Wang et al., 2009) between the surfaces of the resulting segmentation and the corresponding reference segmentation as additional metrics to evaluate the segmentation results. MAD is calculated by measuring the average distance from all points on the surface of the automatic segmentations to the surface of the reference one. In contrast, the symmetric Hausdorff distance computes the maximal distance between the surfaces. The smaller the MAD or Hausdorff distance, the better aligned the points on the two surfaces and thus the better the agreement with the reference segmentation. (Wang et al. 2014)

2.5 Shape Analysis

While the volumetric and overall surface analysis is enough to identify global differences between the automatic and manually edited segmentations, we also wanted to know if there were specific areas, which consistently need manual editing, and to what degree that is necessary and applied structural shape analysis for that purpose. We chose to focus on only the hippocampus because the amygdala has a blobby, almond like shape that does not lend itself well to local shape analysis. We performed the local hippocampal shape analysis via the SPHARM-PDM (Spherical Harmonics Point Distribution Models) analysis toolbox (Styner et al, 2006). SPHARM-PDM allows us to discover the local hippocampal areas that were consistently affected by the manual correction, across both raters and subjects by computing shape statistics at corresponding location along the hippocampal surface.

After computation of corresponding surfaces with SPHARM-PDM, local hippocampal difference vector maps were computed between the hippocampal surface models from the automatic and manually edited segmentations. By projecting the local difference vectors to the local normal at the overall mean surface, the vectors were converted to scalar signed difference maps (Styner et al., 2005). Point-wise paired t-tests using MATLAB (MathWorks) were employed to calculate the statistical significance of the local signed difference at each location independently, resulting in raw significance maps.

3. Results

3.1 Evaluation of Manual Segmentation

3.1.1 Reliability of Manual Segmentation

We computed the intra-class and inter-class correlation for the hippocampal and amygdala segmentations as seen in Table 1. Both within and across raters, we have reliable manual editing of the automatic segmentations. While the inter-class correlations were lower/worse than the intra-class correlations for the hippocampi, the inter CC and intra CC were close to 1 for the amygdalae. The inter-class and intra-class correlations in the left and right amygdalae were practically the same. This surprising fact is due to significant lack of contrast in the amygdala such that little to no edits were completed. Overall, there is limited evidence whether the amygdala is accurate or inaccurate. This result mainly indicates that the MR scans show no evidence contrary to the automatic multi-atlas amygdala segmentation. In contrast the hippocampal boundaries show much stronger contrast and thus were edited more heavily.

Table 1
Intra and inter-class correlations for manual amygdala and hippocampus segmentations.

	Right Hippocampus	Left Hippocampus	Right Amygdala	Left Amygdala
Intra CC	0.994	0.996	0.998	0.999
Inter CC	0.887	0.957	0.998	0.999

Average volumes for right, left hippocampus and right, left amygdala were (1109.97mm^3, 1063.73mm^3) (258.93mm^3, 260.07mm^3). Standard deviations for right, left hippo campus and right, left amygdala were (144.38, 135.57) (31.78, 31.82)

Table 2
Average volume differences between original and mirrored presentation on the segmentations through raw measurement, percent contribution to the overall value and significance of the difference for the hippocampus and amygdala.

	Right Hippocampus	Left Hippocampus	Right Amygdala	Left Amygdala
Raw, mm^3	-1.2	0.2	0.2	0.1
%	-0.10%	0.01%	0.08%	0.03%
P-value	0.070	0.788	0.165	0.558

None of the structures show a significant asymmetric bias due to presentation.

3.1.2 Cases with Significant Edits

Few cases from automatic segmentations required extensive editing. Figure 2 shows the distribution of the manual edits. Both hippocampi and amygdalae were manually edited by a relative volumetric change between 1-4%. Only 1.12% of left hippocampi and 3.37% of left amygdalae needed more manual editing than 4% volumetric change.

3.1.3 Asymmetric Presentation Bias of Manual Segmentation

The average difference in volume for all structures segmented in standard radiological presentation versus segmented in a left-right mirrored presentation is shown in Table 2. None of the structures showed a presentation related asymmetric bias, though the right hippocampus showed a limited trend ($p = 0.070$). Given the presence of a clear asymmetric presentation bias in purely manual segmentations (Maltbie et al, 2012), this result indicates that manual editing of automatic segmentations can reduce such a presentation bias.

Table 3
Average volume differences of the hippocampus and amygdala between manual and automatic segmentations through raw (mm^3) measurements, percent contribution, and significance of difference (results show manual – automatic).

	Right Hippocampus	Left Hippocampus	Right Amygdala	Left Amygdala
Raw, mm^3	-13.5	-16.8	-4.5	-4.6
%	-1.15%	-1.48%	-1.62%	-1.74%
P-value	p<0.000001	p<0.000001	p<0.000001	p<0.000001

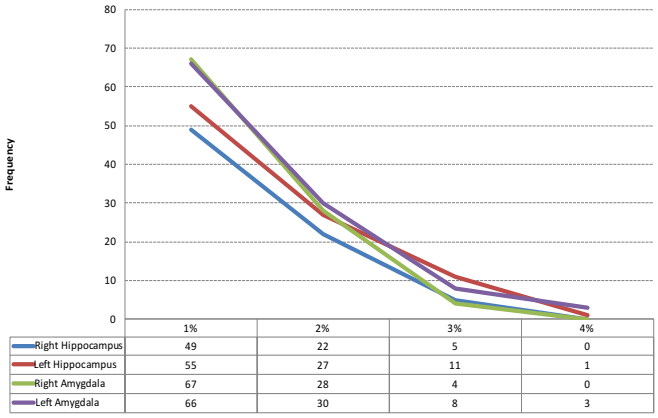
Higher volumetric differences were observed for the left hippocampus than the right hippocampus (p=0.0003), i.e the left hippocampus underwent a significantly higher amount of editing as compared to the right.

Table 4
Coefficients of variation (COV) for the automatic amygdala hippocampus segmentations.

	Right Hippocampus	Left Hippocampus	Right Amygdala	Left Amygdala
COV	0.25%	0.38%	0.22%	0.1%

3.2 Evaluation of Automatic Segmentation

3.2.1 Reliability and Stability of Automatic Segmentation



	1%	2%	3%	4%
Right Hippocampus	49	22	5	0
Left Hippocampus	55	27	11	1
Right Amygdala	67	28	4	0
Left Amygdala	66	30	8	3

Fig. 2. Significant edits of automatic segmentations of the amygdala and hippocampus. Degree of significance is ranked by relative volumetric change from 1-4%.

Table 5
Manual versus automatic segmentation evaluation: average overall surface error, average maximum surface error, and average volume overlap.

	Right Hippocampus	Left Hippocampus	Right Amygdala	Left Amygdala
Average Surface Error (mm)	0.024	0.023	0.017	0.015
Maximum Surface Error (mm)	2.251	1.853	1.919	0.933
Overlap/Dice Coefficient (%)	98.852	98.855	99.143	99.067

COV's were computed over the scan/rescan reliability dataset as seen in Table 4. Overall all structures showed a relatively low level of COV indicating a reliable segmentation as compared to fully manual segmentations of amygdala and hippocampus, which are expected to show COV values in the range of 3-6% (Styner et al., 2002). Thus, our automatic segmentations are appropriately reliable even in the neonate setting where the scans are of borderline quality.

3.2.2 Evaluation of Manual Versus Automatic Segmentations

All hippocampal and amygdala segmentations show a significant decrease in volume between manual and automatic segmentations as shown in Table 3. Overall, the automatic segmentation seems to over-segment both hippocampi and amygdalae. In addition, the left hippocampus segmentations showed a significantly higher decrease ($p=0.0003$, 0.33% larger volume change than the right hippocampus) in volume due to manual editing than the right hippocampus when comparing the automatic with the manual segmentation results.

The average volume overlap error (dice coefficient), average overall surface error, and average maximum surface volume error are shown in Table 5 as a way to measure a shape-based, global agreement between the manual and automatic segmentations. Both left and right hippocampi and both left and right amygdalae display an almost complete volumetric overlap, with average surface errors in the 0.02mm range and maximum surface errors close to 2mm. These results indicate a high agreement between the automatic segmentation and the manual edits.

3.3 Shape Analysis

3.3.1 Average Difference

Local average difference magnitude maps are visualized color-coded on the mean hippocampus surface as seen in Figure 3 (detailed maps from different viewpoints in Supplement Figure 1). Areas where the manual correction reduced the automatic segmentations are shown in orange and areas where the manual correction enlarged the automatic segmentation are color-coded in blue. Unchanged areas are shown in white.

Automatic segmentations needed consistent editing in the head of the left and right hippocampi in the superior CA3 and CA1 regions, where the maps show a large, significant reduction (raw $p < 0.05$) due to the manual correction as seen in Figure 4. There are also clusters of significant volume reductions after manual correction of the automatic segmentation in the superior subiculum of both left and right hippocampi. Consistent editing is found in the inferior subiculum region such that the left and right hippocampus show areas of large, significant reduction after manual correction as well.

Manual corrections in the superior subiculum of the left and right hippocampi resulted in larger volumes in these regions. Visually, the right hippocampus appears to show overall greater enlargement after manual correction in the superior subiculum around the body and tail as well as in the posterior subiculum in the head and body compared to the left hippocampus.

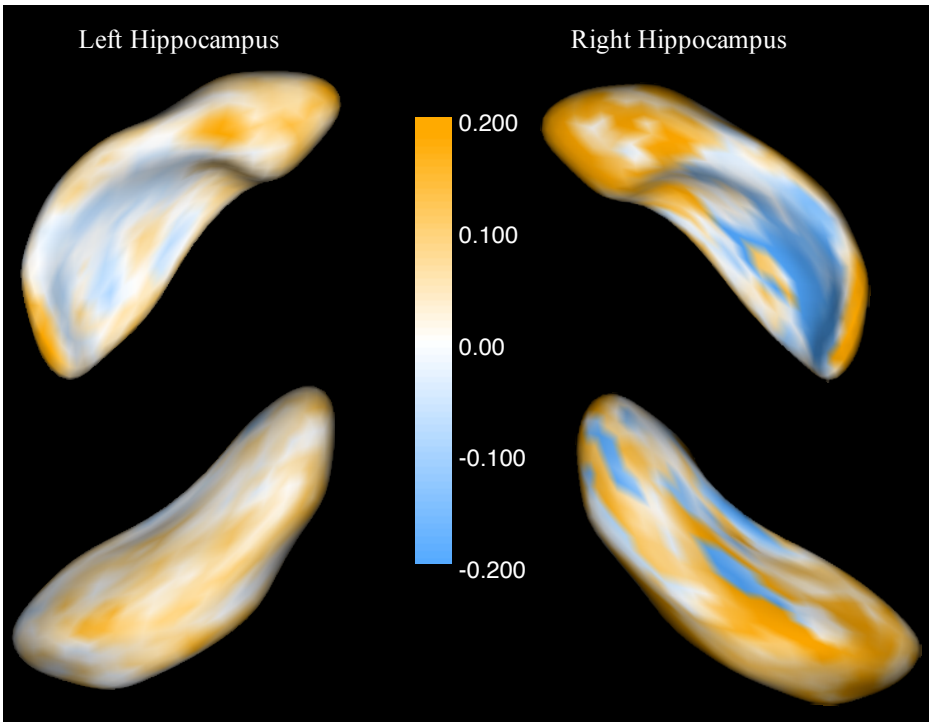


Figure 3. Local average difference magnitude maps (mm) between manual and automatic segmentations of the left and right hippocampus. Areas where the manual correction reduced the automatic segmentation are shown in orange and areas where manual correction enlarged. Top row is visualized from a superior-medial position and bottom row is visualized from an inferior-lateral position.

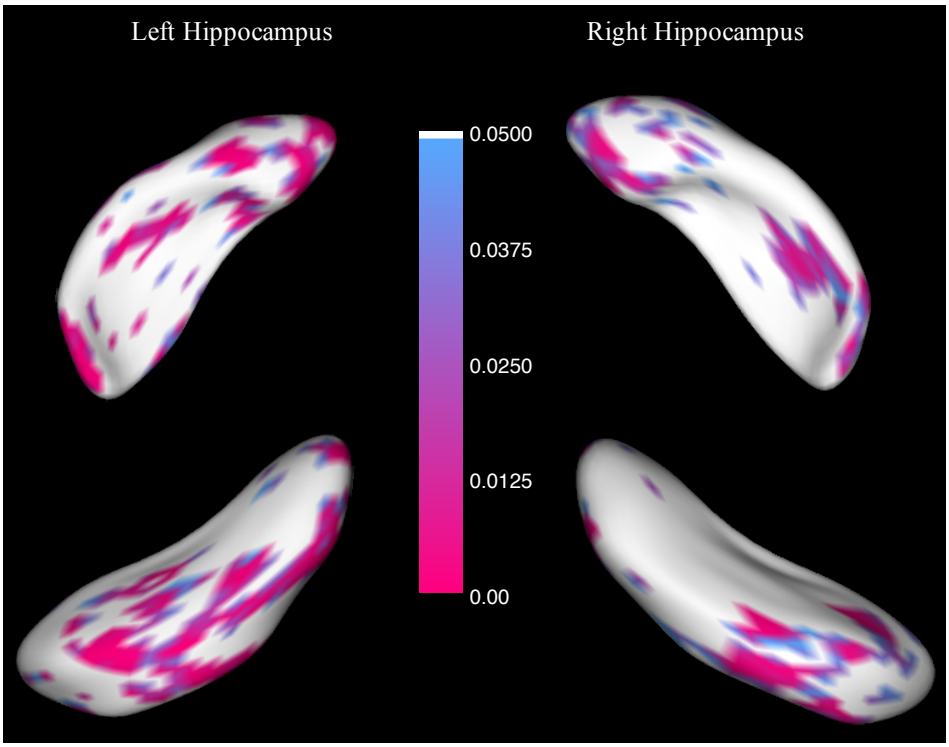


Figure 4. Significance maps for editing changes (raw $p < 0.05$) of the automatic segmentations through manual correction of the left and right hippocampus via paired T-tests. Top row is visualized from a superior-medial position and bottom row is visualized from an inferior-lateral position.

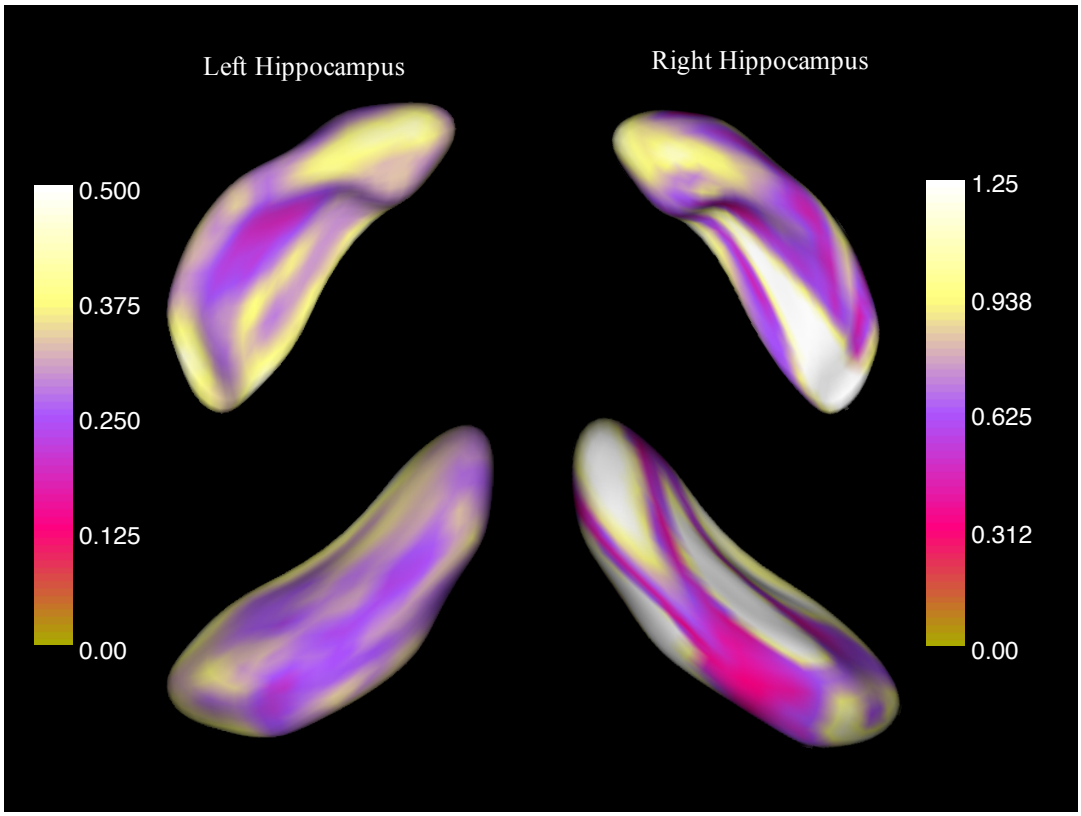


Figure 5. Standard deviation differences (scale in mm) visualized via color-coded maps over the mean lateral hippocampus surface. The left hippocampus STD overall variability is from (0-0.50) while the right hippocampus STD variability is from (0-1.25).

3.3.2 Standard Deviation Difference

Standard deviation differences are presented via color-coded maps visualized over the mean lateral hippocampus surface as seen in Figure 5. The left hippocampus overall showed significantly less variability than the right hippocampus, to a quite surprising degree. The manual editing is overall quite variable even on the left hippocampus with larger areas close or above 0.5mm standard deviation, and similar areas on the right hippocampus displaying standard deviations above 1.5mm. Particularly areas close to the hippocampus-amygdala boundary and the tail section show high variability in edits.

4. Discussion & Conclusion

4.1 Manual Segmentations

Our manual editing of the automatic segmentations of both hippocampus and amygdala showed to be reliable and consistent. It is important to keep in mind that the manual segmentations were performed starting from the automatic segmentation and not from scratch. Amygdalae segmentations for the most part remained untouched as limited sections of the amygdala boundary are clearly visible on these neonate MRI scans. Its contrast is significantly lower than in the hippocampus. This also means that the manual amygdala segmentations are more biased towards the automatic initialization than their hippocampal counterparts. Since the hippocampus has more contrast, it allows the rater to edit the segmentation more extensively while still being reliable.

Given the findings in our prior research (Maltbie et al, 2012) of the presence of widespread asymmetric presentation biases in expert rater segmentations, we were surprised to find no significant evidence of such a bias in our manual editing process of either structure. The amount of manual editing needed here seems to be too limited in order for the method to be sensitive to this bias.

4.2 Automatic Segmentations

Our automatic segmentations for all structures show to be reliable even in scans considered borderline in the neonate setting. But, we find that the automatic segmentation seems to consistently, slightly over-segment both hippocampi and amygdalae.

We found good localized agreement between the manual segmentations and automatic segmentations. We find that the left and right hippocampus and amygdale structures display an almost complete volumetric overlap and low surface errors. For the hippocampus this indicates that the automatic segmentations can be considered anatomically appropriate in the low contrast neonate setting. As mentioned before, the low error for the amygdala segmentations may not necessarily mean that our automatic segmentation yields anatomically accurate amygdala. Amygdala segmentations in neonate MR images are extremely difficult due to the poor tissue contrast and the present results indicate that segmentation cannot be significantly further improved by manual editing. It is important to note though that there is no evidence for our automatic multi-atlas amygdala segmentation being inaccurate.

Only few cases required extensive editing following automatic segmentation. While we did not detect any significant asymmetric presentation bias, the left hippocampus undergoes a higher degree of editing than the right hippocampus. This indicates that while our automatic segmentation method is symmetric (due to the use of original and mirrored presentation atlas images) the results of the methods are slightly, but significantly more accurate on the right than on the left hippocampus.

4.3 Shape Analysis

Several areas in the hippocampus needed consistent editing following the automatic segmentation process. Surprisingly, given the higher degree of editing, the left hippocampus seems to have less editing variability across subjects than the right with respect to the editing. While overall more regions were over-estimated by the automatic segmentation, a few were also consistently under-estimated. This seems to indicate that a further automatic post-processing could further improve the segmentation, e.g. via the trained correction proposed by Wang et al. 2010.

4.4 Conclusion

Here, we show that an automatic segmentation process for the amygdala and hippocampus that is reliable and works well with neonatal MRI scans. We find that manual post-correction is not strictly necessary for appropriate anatomical accuracy of the hippocampus and amygdala. The findings on the amygdalae segmentations must be approached with some caution, both for the automatic and manual results, due to the lack of boundary contrast. We consider the results on the hippocampus to be strong. Our analysis also shows that while manual post-processing is not strictly necessary, it provides consistent differences in regions close to the neighboring regions of the amygdala and hippocampus and thus is expected to increase the accuracy of volume estimates further.

The results presented here were all computed using the AutoSeg platform with the UNC-UCI Hippocampus-Amygdala multi-atlas database. Alongside the UNC/UCI scan/rescan database employed in this work, these resources are all publically available on NITRC.

5. Acknowledgements

We would like to acknowledge the following funding sources MH086633, P50 MH064065, MH070890, HD053000, MH091351, MH091645, HD079124, P50 MH100034, P50 MH100031.

6. Referenced Resources

3D Slicer: http://www.slicer.org
AutoSeg: http://www.nitrc.org/projects/autoseg
itk-SNAP: http://www.itksnap.org/pmwiki/pmwiki.php
MATLAB: http://www.mathworks.com/products/matlab
SPHARM-PDM: https://www.nitrc.org/projects/spharm-pdm
ABC: https://www.nitrc.org/projects/abc/
Hippocampus/Amygdala multi-atlas data & scan/rescan neonate reliability data: https://www.nitrc.org/projects/unc_brain_atlas/

7. References

Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., & Rueckert, D. (2009). Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage*, *46*(3), 726-738.

Bookstein, F. (1991.). *Morphometric tools for landmark data: Geometry and biology.* Cambridge University Press.

Brechbuhler, C. (1995). Description and analysis of 3-D shapes by parametrization of closed surfaces. *Diss., IKT/BIWI, ETH Zurich.*

Brechbuhler, C., Gerig, G., & Kubler, O. (1995). Parametrization of closed surfaces for 3-D shape description. *Comp. Vision, Graphics, and Image Proc., 61*, 151-178. 1, 2.1, 2.2.

Buss, C., Davis, E. P., Shahbaba, B., Pruessner, J. C., Head, K., & Sandman, C. A. (2012). Maternal cortisol over the course of pregnancy and subsequent child amygdala and hippocampus volumes and affective problems. *Proceedings of the National Academy of Sciences*, *109*(20), E1312-E1319.

Chupin, M., Mukuna-Bantumbakulu, A. R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnéhun, S., Lemieux, L., Dubois, B., & Garnero, L. (2007). Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease. *Neuroimage*, *34*(3), 996-1019.

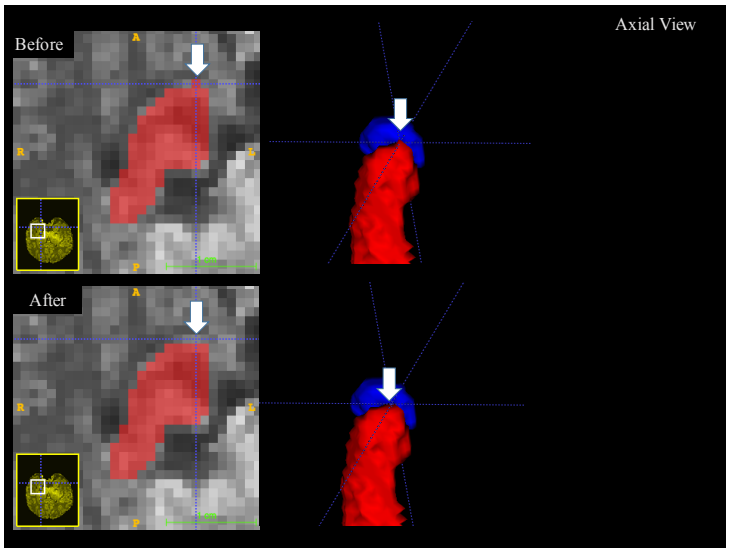
Collins, D. L., & Pruessner, J. C. (2010). Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *Neuroimage*, *52*(4), 1355-1366.

Dager, S. R., Wang, L., Friedman, S. D., Shaw, D. W., Constantino, J. N., Artru, A. A., Dawson, G., & Csernansky, J. G. (2007). Shape mapping of the hippocampus in young children with autism spectrum disorder. *American journal of neuroradiology*, *28*(4), 672-677.

Dill, V., Franco, A. R., & Pinho, M. S. (2014). Automated Methods for Hippocampus Segmentation: the Evolution and a Review of the State of the Art.*Neuroinformatics*, 1-18.

Frodl, T., Stauber, J., Schaaff, N., Koutsouleris, N., Scheuerecker, J., Ewers, M., Omerovic, M., Opgen-Rhein, M., Möllen, H. J., & Meisenzahl, E. (2010). Amygdala reduction in patients with ADHD compared with major depression and healthy volunteers. *Acta Psychiatrica Scandinavica*, *121*(2), 111-118.

Gholipour, A., Akhondi-Asl, A., Estroff, J. A., & Warfield, S. K. (2012). Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly. *Neuroimage*, *60*(3), 1819-1831.



J. H. Gilmore, F. Shi, S. L. Woolson, R. C. Knickmeyer, S. J. Short, W. Lin, H. ZHU, R. M. Hamer, M. Styner, and D. Shen. (2012) Longitudinal development of cortical and subcortical gray matter from birth to 2 years. *Cerebral cortex (New York, N.Y. : 1991)*, vol. 22, no. 11, pp. 2478–2485.

Hajek, T., Kopecek, M., Kozeny, J., Gunde, E., Alda, M., & Höschl, C. (2009). Amygdala volumes in mood disorders—meta-analysis of magnetic resonance volumetry studies. *Journal of affective disorders*, *115*(3), 395-410.

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.,Yushkevich, P., Essex, M., Pollak, S.D., & Davidson, R. J. (2012). Robust automated amygdala segmentation via multi-atlas diffeomorphic registration. *Frontiers in neuroscience*, *6*.

Hu, S., Coupé, P., Pruessner, J. C., & Collins, D. L. (2011). Appearance-based modeling for segmentation of hippocampus and amygdala using multi-contrast MR imaging. *NeuroImage*, *58*(2), 549-559.

Inglese, P., Amoroso, N., Boccardi, M., Bocchetta, M., Bruno, S., Chincarini, A., Errico, R., Frisoni, G. B., Maglietta, R., Ridolfi, A., Sensi, F., Tangaro, S., Tateo, A., & Bellotti, R. (2015). Multiple RF classifier for the hippocampus segmentation: Method and validation on EADC-ADNI Harmonized Hippocampal Protocol. *Physica medica: PM: an international journal devoted to the applications of physics to medicine and biology: official journal of the Italian Association of Biomedical Physics (AIFB)*.

Kesler, S. R., Garrett, A., Bender, B., Yankowitz, J., Zeng, S. M., & Reiss, A. L. (2004). Amygdala and hippocampal volumes in Turner syndrome: a high-resolution MRI study of X-monosomy. *Neuropsychologia*, *42*(14), 1971-1978.

Lehmann, M., Douiri, A., Kim, L. G., Modat, M., Chan, D., Ourselin, S., Barnes, J., & Fox, N. C. (2010). Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements. *Neuroimage*,*49*(3), 2264-2274.

Lötjönen, J. M., Wolz, R., Koikkalainen, J. R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., & Alzheimer's Disease Neuroimaging Initiative. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage*, *49*(3), 2352-2365.

E. Maltbie, K. Bhatt, B. Paniagua, R. G. Smith, M. M. Graves, M. W. Mosconi, S. Peterson, S. White, J. Blocher, M. El-Sayed, H. C. Hazlett, and M. Styner, (2012). Asymmetric bias in user guided segmentations of brain structures., *NeuroImage*, vol. 59, no. 2, pp. 1315–1323.

Morey, R. A., Petty, C. M., Xu, Y., Hayes, J. P., Wagner, H. R., Lewis, D. V., ... & McCarthy, G. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *Neuroimage*, *45*(3), 855-866.

Paniagua, B., Lyall, A., Berger, J. B., Vachet, C., Hamer, R. M., Woolson, S., Lin, W., Gilmore, J., & Styner, M. (2013, March). Lateral ventricle morphology analysis via mean latitude axis. In *SPIE Medical Imaging* (pp. 86720M-86720M). International Society for Optics and Photonics.

Pinkhardt, E. H., van Elst, L. T., Ludolph, A. C., & Kassubek, J. (2006). Amygdala size in amyotrophic lateral sclerosis without dementia: an in vivo study using MRI volumetry. *BMC neurology*, *6*(1), 48.

Rohlfing, T., Russakoff, D. B., & Maurer, C. R. (2003). Expectation maximization strategies for multi-atlas multi-label segmentation. In *Information Processing in Medical Imaging* (pp. 210-221). Springer Berlin Heidelberg.

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

Styner, M. A., Charles, H. C., Park, J., & Gerig, G. (2002). Multisite validation of image analysis methods: assessing intra-and intersite variability. In*Medical Imaging 2002* (pp. 278-286). International Society for Optics and Photonics.

Styner, M., Oguz, I., Xu, S., Brechbuhler, C., Levitt, J., Shenton, M., et al. (2006). Framework for the statistical shape analysis of brain structures using SPHARM-PDM. *Insight Journal, 1071*, 242-250.

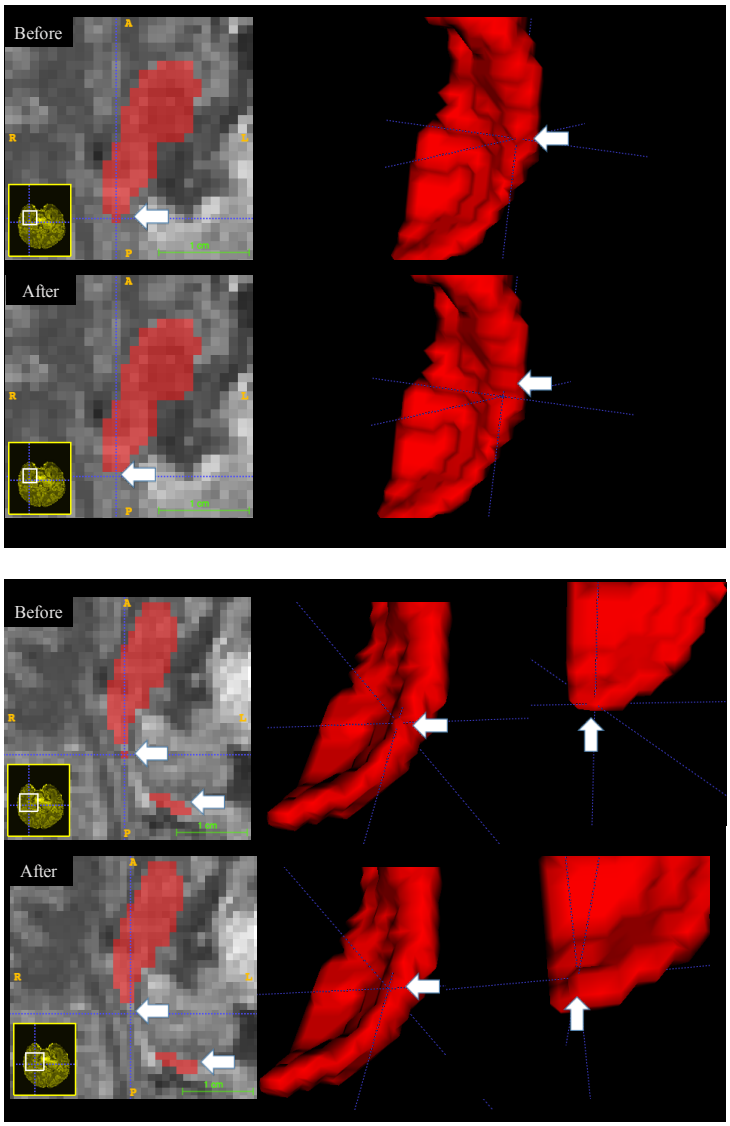
Thompson, D. K., Wood, S. J., Doyle, L. W., Warfield, S. K., Lodygensky, G. A., Anderson, P. J., Egan, G. F., & Inder, T. E. (2008). Neonate hippocampal volumes: Prematurity, perinatal predictors, and 2-year outcome. *Annals of neurology*,*63*(5), 642-651.

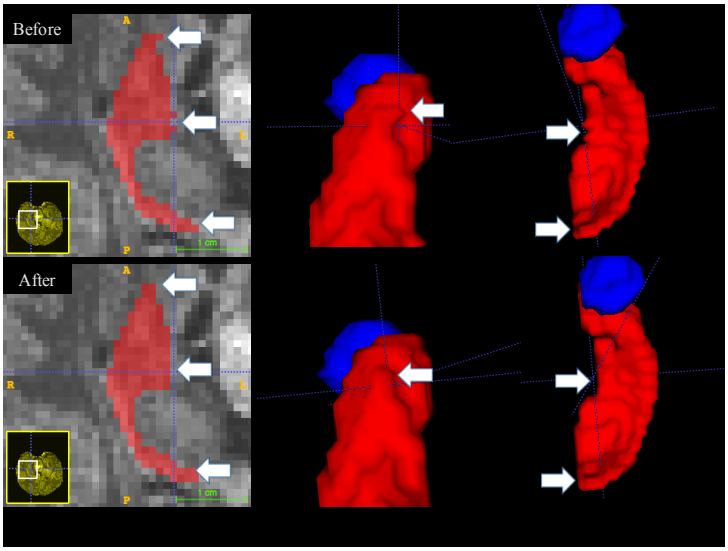
N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, (2010). N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*, vol. 29, no. 6, pp. 1310–1320.

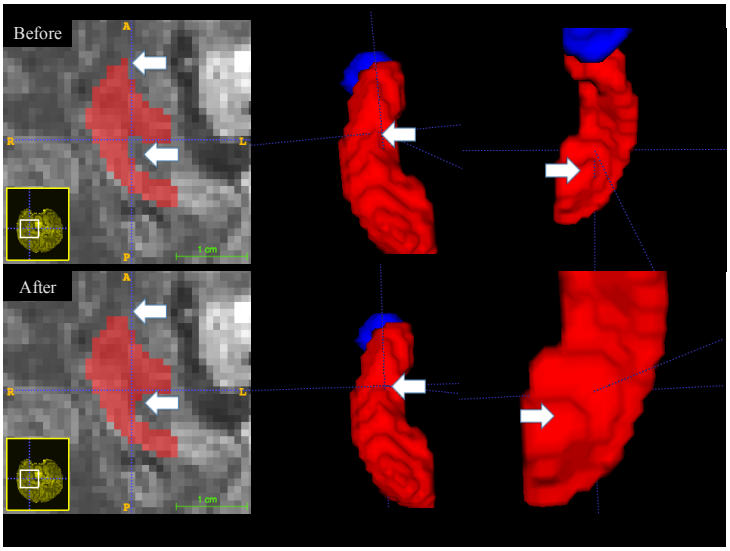
Wang, H., Das, S., Pluta, J., Craige, C., Altinay, M., Avants, B., Weiner, M., Mueller, S., & Yushkevich, P. (2010). Standing on the shoulders of giants: improving medical image segmentation via bias correction. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010* (pp. 105-112). Springer Berlin Heidelberg.

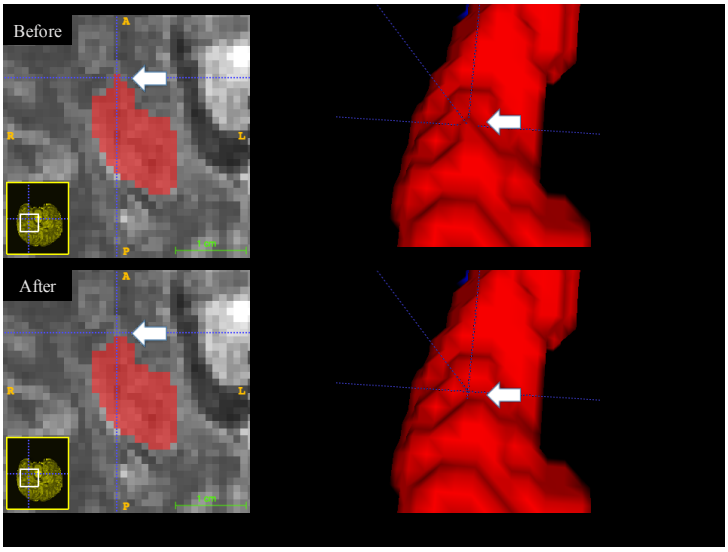
Wang, J., Vachet, C., Rumple, A., Gouttard, S., Ouziel, C., Perrot, E., & Styner, M. (2014). Multi-atlas segmentation of subcortical brain structures via the AutoSeg software pipeline. *Frontiers in Neuroinformatics*, *8*.

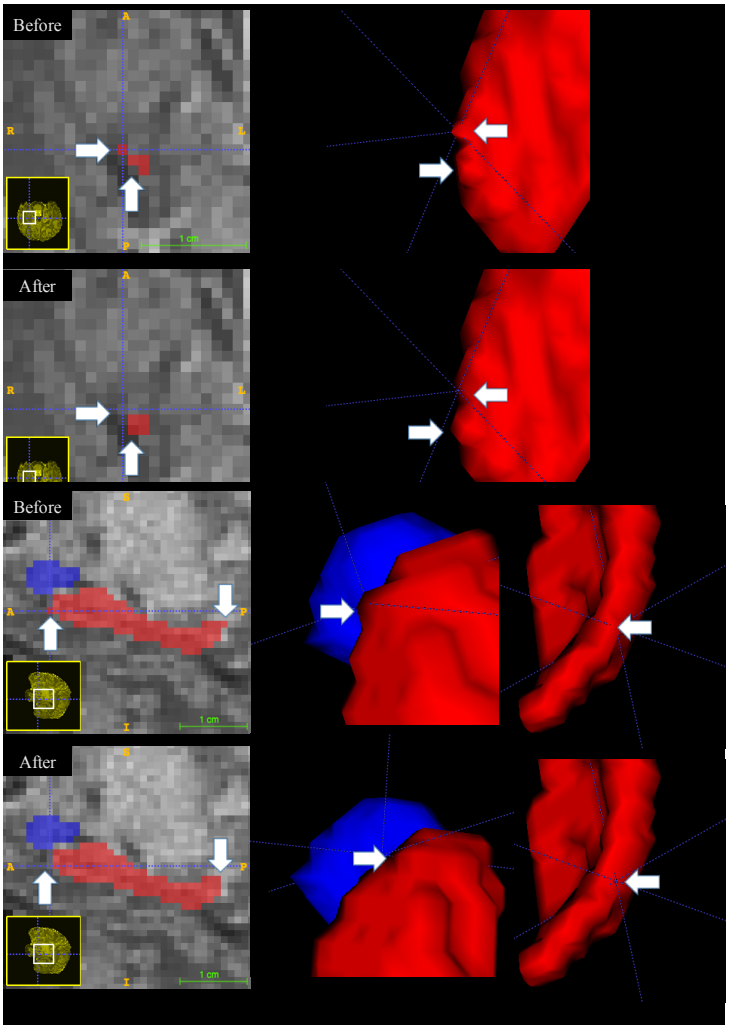
Figure 1. In the axial and saggital views, the rater performs manual correction through each slice. When a voxel appears to be outside of the boundary within the 2D view, the rater checks on the boundary in the 3D view. Once the rater confirms that the voxel is outside the boundary of the hippocampus, the rater manually removes the voxel from the automatic segmentation. When a voxel appears to be missing inside the boundary within the 2D view, the rater looks additionally for the missing voxel in the 3D view. Once the rater confirms that the addition of a voxel is needed within the boundary, the rater adds the voxel to the automatic segmentation.

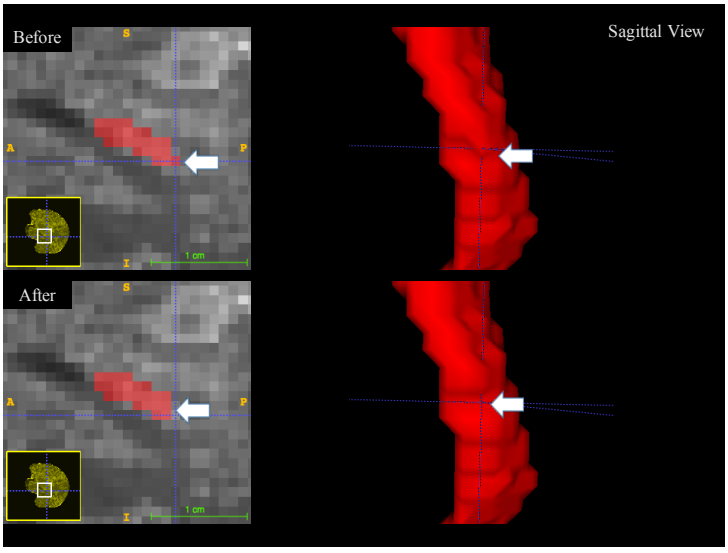


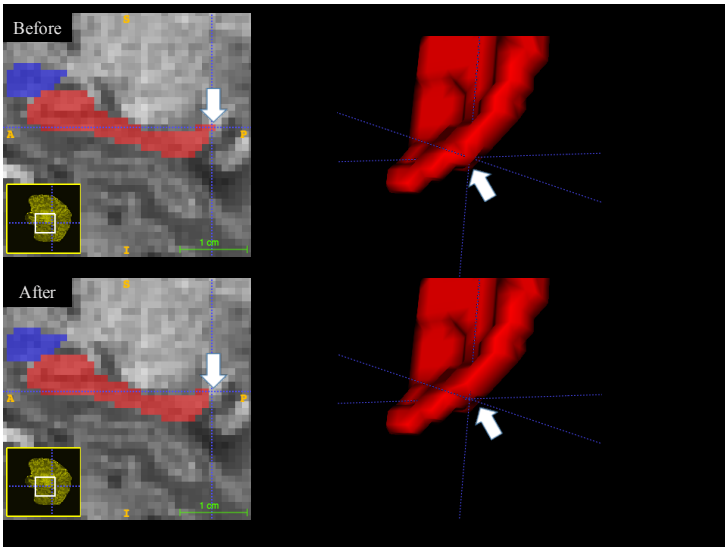

Before
After

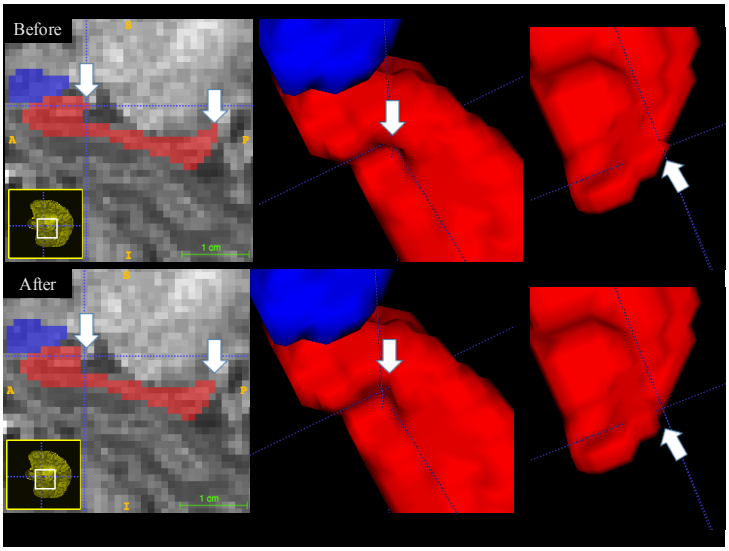

Before
After

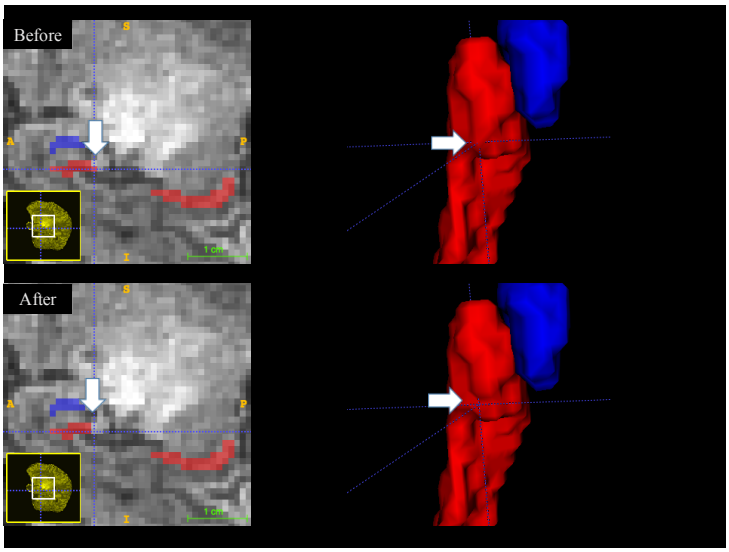

Before
After


Before
After
Before
After


Before
After
Sagittal View


Before
After


Before
After


Before
After

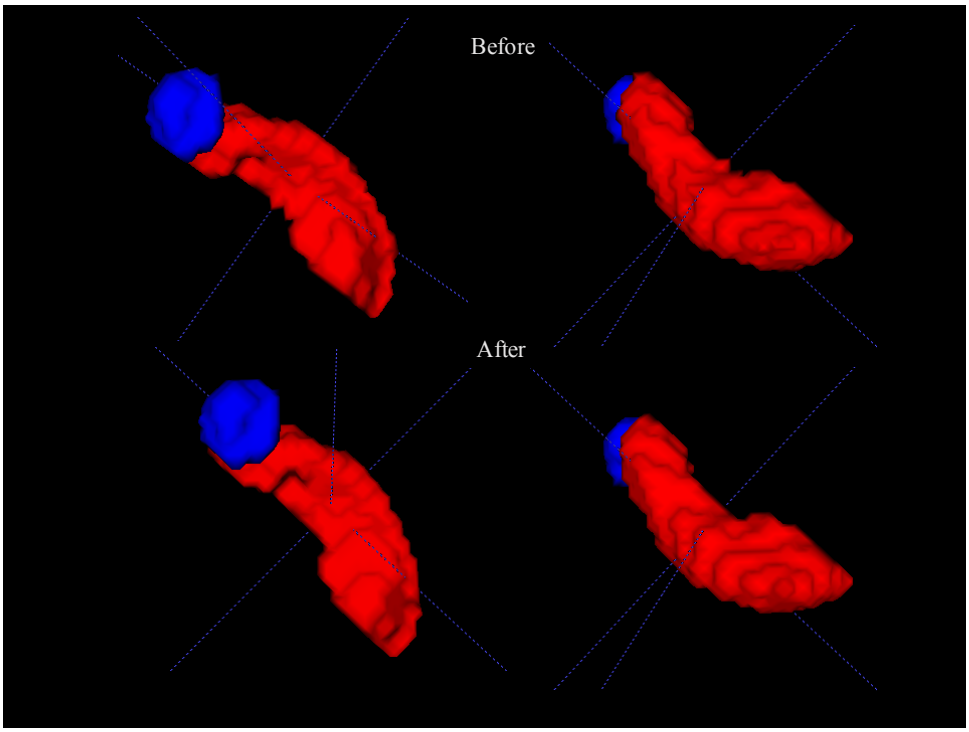


Figure 2. Above shows the automatic segmentation before manual correction. Manual correction by the rater of the right hippocampus is shown below.

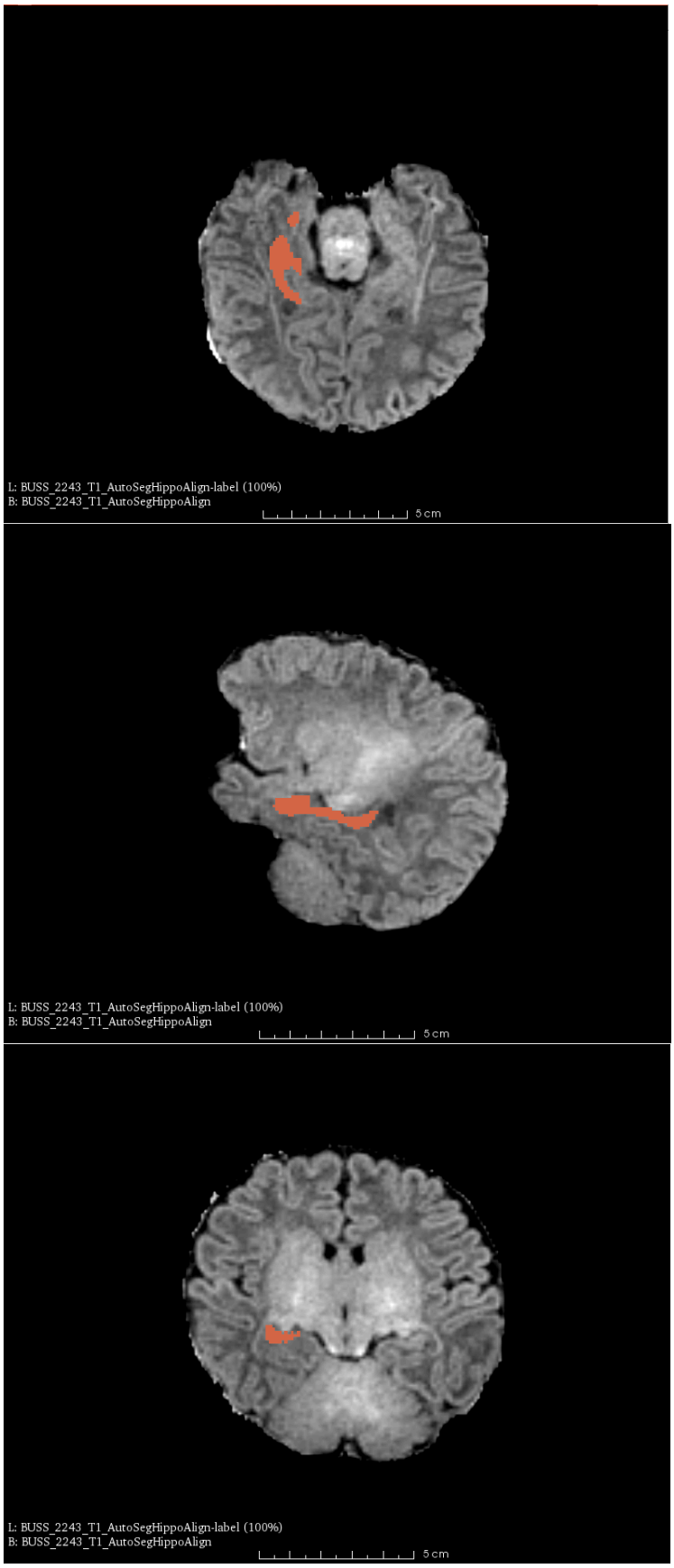


Figure 3. Right hippocampus boundaries overlaying T1-weighted neonate stripped skull brain scan.