HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2020 May 22.

Published in final edited form as:

Neuroimage. 2016 April 01; 129: 1–14. doi:10.1016/j.neuroimage.2016.01.038.

Hippocampus and amygdala volumes from Magnetic Resonance Images in children: assessing accuracy of Freesurfer and FSL against manual segmentation

Dorothee Schoemaker[a,b], **Claudia Buss**[c,d], **Kevin Head**[c], **Curt A. Sandman**[c], **Elysia P Davis**[c,e], **Mallar M Chakravarty**[b,f], **Serge Gauthier**[a], **Jens Pruessner**[a,b]

[a]McGill Centre for Studies in Aging, McGill University, Montreal, QC, Canada

[b]Douglas Hospital Research Centre, Psychiatry Department, McGill University, Montreal, QC, Canada

[c]University of California at Irvine, California, USA

[d]Charité, Berlin, Germany

[e]University of Denver, Colorado, USA

[f]Biomedical Engineering Department, McGill University, Montreal, QC, Canada

Abstract

The volumetric quantification of brain structures is of great interest in pediatric populations because it allows the investigation of different factors influencing neurodevelopment. FreeSurfer and FSL both provide frequently used packages for automatic segmentation of brain structures. In this study, we examined the accuracy and consistency of those two automated protocols relative to manual segmentation, commonly considered as the “gold standard” technique, for estimating hippocampus and amygdala volumes in a sample of preadolescent children aged between 6 to 11 years. The volumes obtained with FreeSurfer and FSL-FIRST were evaluated and compared with manual segmentations with respect to volume difference, spatial agreement and between- and within-method correlations.

Results highlighted a tendency for both automated techniques to overestimate hippocampus and amygdala volumes, in comparison to manual segmentation. This was more pronounced when using FreeSurfer than FSL-FIRST and, for both techniques, the overestimation was more marked for the amygdala than the hippocampus. Pearson correlations support moderate associations between manual tracing and FreeSurfer for hippocampus (right $r=0.69$, $p<0.001$; left $r=0.77$, $p<0.001$) and amygdala (right $r=0.61$, $p<0.001$; left $r=0.67$, $p<0.001$) volumes. Correlation coefficients between manual segmentation and FSL-FIRST were statistically significant (right hippocampus $r=0.59$, $p<0.001$; left hippocampus $r=0.51$, $p<0.001$; right amygdala $r=0.35$,

Address correspondence to: Dr. Jens Pruessner, McGill Centre for Studies in Aging, 6825 Boulevard LaSalle Montreal, QC, H4H1R3, Canada .

p<0.001; left amygdala r=0.31, p<0.001) but were significantly weaker, for all investigated structures. When computing intraclass correlation coefficients between manual tracing and automatic segmentation, all comparisons, except for left hippocampus volume estimated with FreeSurfer, failed to reach 0.70. When looking at each method separately, correlations between left and right hemispheric volumes showed strong associations between bilateral hippocampus and bilateral amygdala volumes when assessed using manual segmentation or FreeSurfer. These correlations were significantly weaker when volumes were assessed with FSL-FIRST. Finally, Bland-Altman plots suggest that the difference between manual and automatic segmentation might be influenced by the volume of the structure, because smaller volumes were associated with larger volume differences between techniques.

These results demonstrate that, at least in a pediatric population, the agreement between amygdala and hippocampus volumes obtained with automated FSL-FIRST and FreeSurfer protocols and those obtained with manual segmentation is not strong. Visual inspection by an informed individual and, if necessary, manual correction of automated segmentation outputs are important to ensure validity of volumetric results and interpretation of related findings.

Keywords

segmentation techniques; pediatric population; hippocampus; amygdala; FSL-FIRST; FreeSurfer

1. Introduction

Childhood is a period of great relevance in the development of risk factors for various neuropsychiatric conditions (Paus et al., 2008). Together with increased efforts in prevention, many large-scale longitudinal studies, starting in early childhood, are currently being undertaken to reveal the impact of environmental, behavioral and biological factors on subsequent developmental outcomes (Chakravarty et al., 2014; Giedd et al., 2015; Raznahan et al., 2014). Due to rapid advances of in-vivo brain imaging technologies, volumetric quantification of brain structures from structural Magnetic Resonance Imaging (MRI) is more accessible than ever. Thus, large-scale studies often acquire MRI to investigate relations between volume of specific brain structures and different aspects of behavior.

Due to their involvement in multiple neuropsychiatric and neurological conditions, the medial temporal lobe structures hippocampus and amygdala have received a considerable amount of attention. The hippocampus is one of the most commonly studied and cited brain structures in the scientific literature. Its involvement in basic cognitive functions, such as memory consolidation (Squire, 1992), psychopathologies such as PTSD (Bonne et al., 2001), major depression (Campbell and MacQueen, 2004), and neurological disorders, such as Alzheimer disease (Fox et al., 1996), is well established. The amygdala is the main structure of the limbic system associated with fear (Adolphs et al., 1994; Davis and Whalen, 2001). It has been linked to many psychopathologies including borderline personality disorder (Donegan et al., 2003; Herpertz et al., 2001), PTSD (Rauch et al., 2000) and social phobia (Stein et al., 2002). The association between negative life events during childhood, such as abuse and traumatic experiences, and the increased risk of developing psychiatric disorders later in life is well documented (Janssen et al., 2004; Johnson et al., 1999;

MacMillan et al., 2001; Springer et al., 2007). It has been hypothesized that the relations between severe childhood stressors and vulnerability to psychopathologies might be mediated trough an impaired development of the hippocampus and/or amygdala (Pynoos et al., 1999; Teicher et al., 2003; Woon and Hedges, 2008). Thus, many efforts are directed at defining and clarifying the roles of the amygdala and the hippocampus in paediatric samples. From a structural neuroimaging perspective, an important challenge lies in the reliable and valid volumetric quantification of these brain regions. However, reliable volumetric estimation is methodologically limited by the anatomical complexity of these two structures.

Manual segmentation is currently considered the gold standard for volumetric quantification of brain structures (Pardoe et al., 2009; Rodionov et al., 2009). However, this procedure requires sufficient anatomical and MR methodological expertise, is difficult and time-consuming to learn, and can be associated with intra- and inter-rater variability if not performed using a consistent approach (Jack Jr et al., 1995). In order to increase reliability and reduce potential biases associated with manual segmentation procedures, multiple protocols have been established and described in the literature for specific target regions (Jack et al., 1990; Matsuoka et al., 2003; Pruessner et al., 2000; Watson et al., 1992). Studies have demonstrated that using these protocols significantly improve intra- and inter-rater agreement (Jack et al., 1990; Matsuoka et al., 2003; Pruessner et al., 2000; Watson et al., 1992). However, these protocols require a considerable amount of training and thus further increase time demands of manual segmentation procedures. In contrast, protocols that offer the fully automated processing and segmentation of target structures from MR images are fast (speed is only limited by CPU power and availability), have excellent reproducibility and require little anatomical expertise from the end user. As a result, a number of automated protocols have recently been developed, published and received favorably by the research community. In part because they are easily and freely accessible to the research community and provide detailed documentation on usage, two of these automated procedures have gained a considerable amount of popularity. The first one is FreeSurfer (http://surfer.nmr.mgh.harvard.edu), a software developed by the Martinos Center for Biomedical Imaging (Fischl et al., 2002). FreeSurfer automatically assigns a label to each voxel from the anatomical image based on probabilistic estimations relying on Markov random fields (MRFs). The localisation and spatial relations between structures are defined according using a training set of manually labelled brains. The second commonly used automated segmentation protocol is “FIRST”, provided as part of the FSL software library (http://fsl.fmrib.ox.ac.uk) (Patenaude, 2007; Patenaude et al., 2011). Using a probabilistic framework, this software estimates boundaries of brain structures based on the signal intensity of the T1 image as well as the expected shape of structures to be segmented.

It is well known that neuroanatomical variations are found not only in clinical populations, but also when comparing brains of normal individuals (Pruessner et al., 2002). Automated segmentation approaches are based on the questionable assumption that computer algorithms can reliably differentiate and delimitate anatomical regions regardless of inter-individual differences in neuroanatomy, scan quality, image contrast, etc. While we did not find any studies comparing the performance of automated segmentation performed with FSL-FIRST and/or FreeSurfer to manual segmentation in pediatric populations, the validity of these protocols has previously been assessed in healthy adult controls (Cherbuin et al.,

2009; Morey et al., 2009; Patenaude et al., 2011) as well as different clinical populations, such as Alzheimer Disease (Pipitone et al., 2014; Sánchez-Benavides et al., 2010; Shen et al., 2010), mood disorders (Doring et al., 2011; Nugent et al., 2013; Tae et al., 2008), temporal-lobe epilepsy (Akhondi-Asl et al., 2011; Pardoe et al., 2009) and psychosis (Pipitone et al., 2014). These reports generally support the ability of automated methods to detect volume difference between clinical groups. However, many articles have highlighted a tendency for FreeSurfer and FSL-FIRST to overestimate volume of brain structures (Cherbuin et al., 2009; Doring et al., 2011; Morey et al., 2009; Nugent et al., 2013; Pipitone et al., 2014; Sánchez-Benavides et al., 2010; Shen et al., 2010; Tae et al., 2008). When assessing the correspondence between volumes derived from these two automated protocols and manual segmentation earlier findings are variable. For the hippocampus region, results usually support moderate to strong associations between manual tracing and FreeSurfer, with Pearson correlation coefficients ranging from 0.71 (Cherbuin et al., 2009; Sánchez-Benavides et al., 2010) to 0.90 (Shen et al., 2010). Studies looking at the association between hippocampus volumes derived from FSL-FIRST and manual segmentation report Pearson correlations ranging from 0.47 (Pardoe et al., 2009) to 0.67 (Nugent et al., 2013). Few studies have looked at the agreement between amygdala volumes derived from automated segmentation protocols and manual tracing. A study by Morey et al. (2009) revealed weaker associations between manual segmentation and both FSL-FIRST and FreeSurfer when estimating the amygdala volume than when estimating the hippocampus volume (Morey et al., 2009). Taken together, these results seem to indicate that the concordance between volumes derived from manual segmentation versus automatic protocols depend on the segmented structure as well as the protocol used. Further, a report by Sánchez-Benavides suggests that the accuracy of automated protocols may vary depending on neuroanatomical characteristics of studied populations (Sánchez-Benavides et al., 2010). More precisely, this later study highlights a larger discrepancy between manually and automatically segmented volumes when used on atrophic brains. Previous reports assessing the validity and accuracy of FSL-FIRST and FreeSurfer were based on adult brains; it remains uncertain whether smaller brain volumes and potential changes in gray / white matter contrasts in pediatric brains negatively affect the performance of these two automated segmentation software. Thus, studies investigating the validity of automated segmentation in children are needed.

The goal of this article was to explore the validity of FSL-FIRST and FreeSurfer in estimating hippocampus and amygdala volumes in children. To do so, we compared volumes generated by these two automated techniques to volumes obtained by manual segmentation, which is considered to be the “gold standard” approach. The validity of the segmentation methods was investigated by means of three different approaches. First, we established discrepancies between volumes derived from manual segmentation and automated methods. Second, to estimate the consistency between manual and automated segmentation, we assessed between- and within-method associations of hippocampus and amygdala volumes. Finally, to explore agreement between volumes and estimate possible proportional and fixed biases in volume estimation we computed Bland-Altman plots.

2. Methods

2.1 Subjects

Anatomical MRI scans were collected in preadolescent children as part of two studies on child neurodevelopment that applied the same MRI acquisition protocol conducted at the University of California Irvine (Buss et al., 2012; Davis et al., 2013). Institutional review boards from all participating institutions approved all study procedures. All T1 images were visually inspected for quality of the image and for absence of apparent motion artefacts. 153 scans judged to be of good quality were retained and used in this study. Two scans were removed due to co-registration issues when using FSL and 4 others were removed due to inadequate processing with FreeSurfer, leaving 147 subjects for final analyses. Following quality control, the final sample included 65 girls and 82 boys study (age range: 6 to 11 years, mean age = 8.47 years ± 1.37 SD). These children were predominantly right-handed (n=130). The demographic information of subjects used in analyses is summarized in Table 1.

2.2 MRI acquisition

T1 anatomical imaging was performed on a 3-T Philips Achieva MRI scanner, at 1mm isotropic resolution. Images were acquired in the sagittal plane with the following parameters: repetition time 11 ms; echo time 3.3 ms; inversion time 100 ms; turbo field echo factor 192; 150 slices; sensitivity encoding for fast MRI acceleration; and flip angle 18°.

2.3 Volumetric quantification

2.3.1 Manual segmentation of the hippocampus and amygdala—Before proceeding to the manual segmentation, anatomical images were corrected for intensity non-uniformity (Sled et al., 1998) and registered to the stereotaxic space (MNI152 template) (Collins et al., 1994) using a linear transformation. This pre-processing was performed to facilitate the identification of key structures and improve segmentation consistency between scans. The hippocampus and amygdala were segmented by a single expert rater using the software DISPLAY (www.bic.mni.mcgill.ca/software/Display/Display.html). The anatomical borders of the two key structures were defined and segmented according to the protocol described by Pruessner et al. (Pruessner et al., 2000). As the structural characteristics, delineations and boundaries of the hippocampus and amygdala in children aged over 6 years old are fully developed (Arnold and Trojanowski, 1996), the segmentation protocol was used as described in the article and no specific modifications were necessary for the population of interest. This protocol has been shown to allow good intra- and inter-rater reliability. Consistently, the present rater achieved an intraclass correlation coefficient of 0.90, and an intrarater reliability of 0.92. One of the main objectives of this study was to define whether smaller brain volumes could affect the accuracy of FSL-FIRST and/or FreeSurfer in estimating hippocampus and amygdala volumes. Therefore, we used original MR T1 images from children participants as input for both automated protocols. Consequently, to be able to compare all segmentation methods within the same space, labels from manual segmentation were resampled to the native space using the inversion of the matrix file designed to perform the linear transformation prior to the manual segmentation. Native labels from the specific structures (left/right amygdala and hippocampus) were saved

as four distinct binary masks, each representing a single structure. A voxel count was then used to estimate volumes from manually segmented structures. To verify that the resampling of labels did not influence our results and conclusions, we also computed native volumes by dividing the original segmentation volume in standard space by the global scale factor associated with the linear transformation (native volume = standard volume/[x * y *z scale factors]). Volume difference and between-methods correlation analyses described below were also performed with native volumes obtained the using the global scaling factor.

2.3.2 Automated segmentation of the hippocampus and amygdala using FreeSurfer—

The segmentation of the hippocampus and amygdala were also performed using the FreeSurfer "recon-all" pipeline (v.4.4.0; http://surfer.nmr.mgh.harvard.edu/). In brief, this technique estimates the probability of each voxel to belong to a certain structure, based on a-priori knowledge of spatial relationships acquired with a training set. It uses differences in voxel intensity to locate and parcelate subcortical structures and affine registration to the Talairach space. The FreeSurfer processing stages are fully described in Fishl et al. (2002). All files were visually inspected to ensure adequate registration. Four subjects were removed from the analysis due to poor co-registration. The volumes provided in the aseg.stats file were used in the analysis, because these take into account partial volume estimation and are judged to be more accurate than the voxel count of label files. For visualization, segmentation files in the native space were converted into the MINC format. Labels from the specific structures (left/right amygdala and hippocampus) were also saved as four distinct binary masks in the native space.

2.3.3 Automated segmentation of the hippocampus and amygdala using FSL—

Hippocampus and amygdala volumes were further obtained using FSL-FIRST (v.1.2; http://fsl.fmrib.ox.ac.uk/). In brief, following registration to a standard template this software uses a Bayesian probabilistic model that relies on shape and intensity to infer the location of structures of interest. For each structure a pre-defined number of modes is applied to ensure the best fit. More documentation on the processing steps of FIRST can be found in Patenaude's articles (Patenaude, 2007; Patenaude et al., 2011). Finally, segmentation labels in the native space were converted in the MINC file format. All files were visually inspected to ensure correct registration. Two subjects were removed from subsequent analyses due to inadequate co-registration and poor processing. Labels from the specific structures (left/right amygdala and hippocampus) were saved as binary masks, generating four separate masks. A voxel count was then used to estimate volumes of structures segmented using FSL-FIRST.

2.4 Statistical analysis

Volumes used for method comparisons and statistical analyses were in the native space. Due to the absence of group comparisons or correlations with external factors in the current analyses, we did not correct for intracranial volume as there was no specific need to control for this variable. All the following statistical analyses were performed using IBM SPSS statistics version 20.

2.4.1 Analysis of volume difference—

The percentage of difference between volumes obtained with automated methods and manual segmentation was computed using the

following formula: %VD= [(*Va*- *Vm*)/ *Vm*]*100%. In the event that the automated (V_a) method reaches an identical volume as manual segmentation (V_m), the resulting percentage of volume difference (VD) would be 0 %. Hence, larger percentages of VD indicate increased discrepancy between the volume derived from manual segmentation and volumes derived from automated methods. Negative values are indicative of an underestimation of volumes, in comparison with manual segmentation, while positive values suggest an overestimation of volumes computed automatically relative to manual segmentation. In order to investigate potential interactions between methods and segmented area (as expressed in percentages of volume difference), we conducted a two-factor (Method x Area) repeated measure ANOVA. Significant main effects were explored using post hoc Bonferroni-corrected paired-samples t tests with a significance threshold adjusted to p. <0.01 to account for the four (k=4) performed comparisons. To locate regions of disagreement between volumes derived from automatic methods and manual segmentation, 3D maps of regional differences were prepared. For each subject, the transformation matrix associated with registration to the MNI152 space was estimated using the “mritotal” tool of the MINC Tool Kit. Binary masks representing labels from each of the three segmentation methods were then resampled to the MNI152 space, using the same transformation matrix. Using the “mincmath” tool of the MINC Tool Kit maps of regional agreement between manual segmentation and both automated techniques were computed. Specifically, these maps were constructed so that each voxel represents the average percent of volume difference between labels from manual segmentation and the automated method (100%, indicating a total disagreement that a specific voxel belongs to the segmented structure and 0%, indicating a total agreement). Thus, a voxel with a percentage difference value of 25% would indicate that in 25% of the subjects where this specific voxel is inconsistently labelled between techniques, while in 75% of subjects, this voxel is labeled by both techniques. For visualisation, the maps are presented on the average standardized brain of all participants included in the analyses.

2.4.2 Correlation analysis—Pearson correlations were conducted to estimate associations between manual and automated techniques and to establish whether volumes derived from automated methods are significantly associated volumes obtained with manual segmentation. A strong correlation would confirm a good consistency between automated techniques and manual segmentation. To compare the two automated segmentation techniques with regards to their correlation with manual segmentation, we computed Steiger’s *z* test, a test recommended to assess the difference in magnitude between correlated and overlapping correlation coefficients (Meng et al., 1992; Steiger, 1980). Further, to obtain a concurrent estimate of consistency and agreement between volumes derived from the different segmentation techniques, we computed intraclass correlation coefficients (ICC) (Shrout and Fleiss, 1979). An ICC value of 1 indicates a perfect reproducibility between two (or more) raters and of 0 or less, a reproducibility that is lower than what is expected on the basis of chance alone. While there is no official guideline for the interpretation of ICCs, it has previously been suggested that a ICC denoting a good reproducibility between measurements should be equal to or higher than 0.75 (Burdock et al., 1963). Further, 0.70 has often been considered as the minimum standard for adequate reliability (Nunnally et al., 1967; Terwee et al., 2007). ICCs were computed automatically with SPSS and, specifying a

mixed-effect model as per Shrout & Fleiss' (1979) guidelines. Finally, to assess within-method consistency, Pearson correlations were performed between volumes of bilateral structures segmented within a same technique. Past research indicates that, in a single subject, a moderate to strong association is expected between homotopic (left versus right hemisphere) volumes (Allen et al., 2002). Weak left versus right hemisphere correlations would indirectly suggest a lack of consistency or the presence of errors in volume estimation within the assessed method. Further, if the two automated segmentation protocols are consistent with manual segmentation, similar associations between left and right hemisphere volumes are expected when comparing theses methods. Thus difference in magnitude between within-method correlations was also assessed according to the statistical procedure described in Raghunathan et al. (1996) article and based on the Fisher r-to-Z transform (*ZPF*) (Raghunathan et al., 1996). In comparison to the Steiger's z statistical test, this procedure is designed to assess differences between correlated but nonoverlapping correlation coefficients.

2.4.3 Analysis of estimation biases—To further investigate agreement between manual segmentation volumes and volumes derived from automated protocols, we computed Bland-Altman plots. This graphical method is used to illustrate differences in estimation between two techniques or raters (Bland and Altman, 1986). Bland-Altman plots are sometimes created using the mean of the two studied techniques as the estimation of reference. However, as manual segmentation is accepted and viewed as the gold standard of technique for hippocampus and amygdala volumes estimation, we plotted the difference between automated and manually segmentation volumes against the volumes obtained with manual segmentation. Arguments in favor of this procedure can be found in Krouwer et al. 2008 (Krouwer, 2008). We further integrated a regression line to the plot to explore possible biases in volume estimation and observe whether characteristics of studied brain structures, as defined using the gold standard technique, influence the discrepancy between manually and automatically segmented volumes.

3. Results

3.1 Analysis of volume differences

Percentages of volume difference were computed separately for the left and right hippocampus and the left and right amygdala. The mean percentage of volume difference of FreeSurfer-derived volumes relative to manually segmented volumes was of 60.38% (SD=13.04) and 51.53% (SD=13.17) for the left and right hippocampi, respectively, and 100.29% (SD=24.56) and 93.56% (SD=25.78) for the left and right amygdala, respectively. When computing the difference between FSL-FIRST and manual segmentation, the mean percentage of volume difference was of 27.61% (SD=14.49) and 28.39% (SD=13.07) for the left and right hippocampi, respectively and of 50.32% (SD=27.65) and 40.29% (SD=26.09) for the left and right amygdala, respectively. The mean hippocampus and amygdala volumes as well as percentage of volume difference derived from each technique are presented in Table 2. The effects of the segmentation technique (FSL-FIRST versus FreeSurfer) and the segmented area (average left and right hippocampus volume respectively average left and right amygdala) on the obtained percentage of volume difference were tested with a two-way

repeated measure ANOVA. This analysis revealed a significant effect of the technique $F(1,146) = 1555.65$, $p. < 0001$. Post-hoc Bonferroni-corrected pairwise comparisons further revealed that FreeSurfer leads to significantly larger percentage of volume difference than FSL-FIRST for both the hippocampus ($t(146) = 38.24$, $p. < 0001$) and the amydgala ($t(146) = 29.52$, $p. < 0001$). A highly significant effect of the segmented area was also noted ($F(1,146) = 395.22$, $p. < 0001$). Bonferroni- corrected pairwise comparisons showed that the amygdala yielded significantly larger percentage of volume difference than the hippocampus when segmented with both FSL-FIRST ($t(146) = 9.85$, $p. < 0001$) and FreeSurfer ($t(146) = 24.11$, $p. < 0001$). Further, there was a significant interaction effect between the automated segmentation method and the area ($F(1,146) = 180.27$, $p. < 0001$), due to the fact that the difference in volume differences between the hippocampus and the amygdala was even more pronounced when using FreeSurfer than FSL-FIRST. Results of this analysis are summarized in Figure 1. To obtain a visual estimation of areas of discrepancy between manual segmentation and the two studied automated methods, 3D-maps where computed for FSL-FIRST vs. manual segmentation (Fig. 2), and FreeSurfer (Fig 3) vs. manual segmentation using the 'mincmath' command, as part of the Mine ToolKit for manipulating 3D images (http://www.bic.mni.mcgill.ca/ServicesSoftware/MINC). As expected, these maps showed that, while the agreement between manual and automated segmentation is usually satisfactory toward the inner sections of the structures, especially at the cores, the disagreement increases linearly towards the lateral and medial, superior and inferior, and anterior and posterior borders of the target structures. From Figures 2 and 3, it appears that higher percentages of difference appear in the hippocampal tail as compared to the head area. For the amygdala, when comparing FSL-FIRST against manual segmentation, higher percentages of difference are noted in superior boundaries. The same comparison between FreeSurfer and manual segmentation shows differences in both superior and inferior boundaries.

3.2 Correlation analysis

3.2.1 Between-method correlations—Pearson correlations between manual segmentation and FreeSurfer volumes were $r_{rhc} = 0.69$ and $r_{lhc} = 0.77$ for right and left hippocampus, respectively and $r_{rag} = 0.61$ and $ri_{ag} = 0.67$ for right and left amygdala, respectively. Correlations between FSL-FIRST and manually segmented volumes were $r_{rhc} = 0.59$ and $r_{lhc} = 0.51$ for the right and left hippocampus, respectively and $r_{rag} = 0.35$ and $r_{lag} = 0.31$ for the right and left amygdala, respectively. All correlations reached a $p<0.0001$ threshold. Correlations between volumes obtained with manual segmentation and automatic protocols for FreeSurfer and in for FSL-FIRST are summarized in Fig. 4 A and B, respectively. For each region (lhc, lag, rhc, rag), the difference in magnitude between correlations obtained with FSL-FIRST and the one obtained with Freesurfer was tested using the Steiger's z test. Since a total of four comparisons were performed, the alpha was adjusted to $p < .01$ for statistical significance, applying the Bonferroni correction. Correlations between manual and automated segmentation volumes were significantly stronger for Freesurfer than FSL-FIRST for the left ($Z=4.83$, $p<0.001$) and right ($Z=3.31$, $p<0.001$) amygdala and the left hippocampus ($Z=5.05$, $p<0.001$). For the right hippocampus, the difference in correlations obtained with manual segmentation obtained with FSL-FIRST and FreeSurfer did not reach our corrected significance threshold ($Z=2.28$, $p=0.01$). To

investigate causes of incongruity between segmentation volumes, outliers were identified using the magnitude of the residuals and selecting individuals that were at the furthest distance from the regression line. Illustrations of the segmentation obtained from these outliers are presented in Figures 5 and 6 for FSL-First and FreeSurfer, respectively.

3.2.2 Intraclass Correlation Coefficient—The ICC between manual segmentation and Freesurfer was r_{lhc} = 0.74 (CI: 0.66–0.81) for the left hippocampus, r_{rhc} = 0.68 (CI: 0.59–0.76) for the right hippocampus, r_{lag} = 0.65 (CI: 0.55–0.74) for the left amygdala and r_{rag} = 0.60 (CI: 0.48–0.69) for the right amygdala. When comparing manual segmentation and FSL-FIRST volumes, the ICC for the left hippocampus was r_{lhc} = 0.51 (CI: 0.38–0.62), r_{rhc} = 0.59 (CI: 0.47–0.68) for the right hippocampus, r_{lag} = 0.30 (CI: 0.15–0.44) for the left amygdala, and r_{rag} = 0.33 (CI: 0.17–0.46) for the right amygdala.

3.2.3 Within-method correlation analysis—Pearson correlations between volumes in the left and right hemisphere derived from each technique were calculated to estimate within-method consistency. Results of this analysis are presented in Fig. 7 (A to F). The association between interhemispheric (left versus right) volumes was r = 0.85 (p.<0.0001) for hippocampus and r = 0.75 (p. <0.0001) for amygdala volumes estimated with manual segmentation, r = 0.83 (p.<0.0001) for hippocampus and r = 0.77 (p.<0.0001) for amygdala volumes estimated with FreeSurfer, and r = 0.53 (p.<0.0001) for hippocampus and r = 0.59 (p.<0.0001) for amygdala volumes estimated with FSL-FIRST. The difference in magnitude between the computed correlations was tested with the ZPF statistic. Overall, 4 comparisons were performed: correlations between bilateral hippocampi (bHC) volumes estimated with manual segmentation versus correlations between bHC volumes estimated with FSL-FIRST/ FreeSurfer; correlations between bilateral amygdala (bAG) volumes estimated with manual segmentation versus correlations between bAG volumes estimated with FSL-FIRST/ FreeSurfer. Consequently, the alpha was adjusted to $p < .01$ for statistical significance, as per the Bonferroni procedure. Using this criterion, significant differences were observed only between within-method correlations of volumes estimated with manual segmentation and with FSL-FIRST. More precisely, the results suggest a stronger association between bi-hemispheric volumes when estimated with manual segmentation than FSL- FIRST. This was true for both the bAG (ZPF = 2.55, $p < .01$) and bHC (ZPF = 6.21, $p < .01$) volumes. No significant difference was found between the strength of within-method correlations of bHC (ZPF = 0.66, $p > .05$) and bAG (ZPF = −0.58, $p > .05$) volumes when estimated with manual segmentation or with FreeSurfer.

3.3 Analysis of estimation biases

Bland-Altman graphs plotting raw volume difference between manual and automatic segmentation volumes against manual segmentation volume, considered to be the “gold standard” measure, confirm that both FreeSurfer and FSL-FIRST (Fig. 8 A and B) yielded larger volumes than manual segmentation. In all plots but the one comparing left hippocampus volumes between FreeSurfer and manual segmentation, the incorporated regression line highlights a negative linear trend between volume difference and baseline manual segmentation volume. This suggests that smaller volume of the studied structures leads to larger difference in volume estimation when comparing automatic to manual

tracing. Thus, this seems to indicate that neuroanatomical features possibly systematically influence outputs from automatic segmentation protocols.

4. Discussion

Here we compared two widely used automated segmentation tools, FSL-FIRST and FreeSurfer, against manual segmentation, the current gold standard technique, for estimating hippocampus and amygdala volumes in a population of preadolescent children. To our knowledge this is the first study looking at the validity of automated segmentation tools in a large pediatric sample. In this study, we decided to focus on hippocampus and amygdala volumes because these regions are implicated in multiple psychopathologies and are among the most commonly studied in the field of neuroscience. We also defined manual volumes as the standard of reference, because its validity has been established in previous articles (Pardoe et al., 2009; Rodionov et al., 2009).

Our results highlight important differences between volumes derived from manual segmentation and the two studied automated techniques. Indeed, both FreeSurfer and FSL-FIRST overestimated total hippocampus and amygdala volumes in comparison with the manual segmentation protocol used in the current study. When the same volume difference analyses were performed using native volumes obtained by dividing the volume of labels manually segmented in the standard space by scale factors of the linear transformation (x*y*z), the results were highly similar and also suggested that FreeSurfer and FSL-FIRST overestimated hippocampus and amygdala volumes in comparison to manual segmentation. This suggests that large volume differences between manual and automated segmentation were not due to biases associated to the resampling of labels. Further, this tendency for volume overestimation has been reported in earlier work in non-pediatric populations (Cherbuin et al., 2009; Doring et al., 2011; Morey et al., 2009; Nugent et al., 2013; Pipitone et al., 2014; Sánchez-Benavides et al., 2010; Shen et al., 2010; Tae et al., 2008). Between the two automated approaches, FreeSurfer was found to yield the largest volume estimates. Our results further showed that the overestimation of volumes associated with automated segmentation was more pronounced for the amygdala than for the hippocampus. This was true for both automated methods, but was also more pronounced with the FreeSurfer method. To better understand the origin of volumetric overestimation that occur with these automated techniques, 3D neuroanatomical maps representing the average percentage of difference between automatic and manual segmentation were computed to localize areas of disagreement. A qualitative revision of those maps revealed that areas of disagreement were located at the border of the target structures, found in all dimensions (x-y-z axis), rather than in one specific location, or in only one dimension. This suggests that the difference in volumes was likely not a result of differences in the anatomical definition of the target structures, but rather a too liberal inclusion of voxels towards the structure boundaries. This might perhaps be explained by partial volume effects, which can lead to incorrect inclusion of voxels neighbouring the target structure. Thus, it appears likely that automatic segmentation techniques that were tested are more susceptible to partial volume segmentation faults when compared to manual segmentation.

However, it cannot be excluded that differences in volumes obtained between manual segmentation and automated protocols reflect variations in the definition of anatomical boundaries between segmentation protocols. Manual segmentation of the hippocampus and amygdala performed in this study was based on the protocol established by Pruessner et al. in 2000 (Pruessner et al., 2000). FreeSurfer and FSL-FIRST pipelines are based on manual labels provided by the Center for Morphometric Analysis, part of the Massachusetts General Hospital. More details on the segmentation protocols used by this Center can be found at www.cma.mgh.harvard.edu/manuals/segmentation. The protocol used for the manual segmentations in this article systematically excludes the Andreas-Retzius and the Fasciolar gyrus from the tail of the hippocampus. Also, this protocol takes extra care to avoid including the inferior horn of the lateral ventricle, even in subjects where it might not be clearly apparent, by excluding voxels in the infero-lateral portion of the hippocampus with ambiguous signal intensity. This exclusion takes place even if in one slice these voxels appear as gray matter, but the existence of the inferior horn can be extrapolated from neighboring slices. Such an approach is likely not present in automatic segmentation methods for hippocampal volumes, and thus can be expected to result in somewhat larger volume estimates. However, the amount of volume that would be generated by the inclusion of the Andreas-Retzius gyrus and the lateral ventricle can be estimated not to be more than 5% additional volume, which is far inferior to the volume differences observed between the automated methods and the manual one in the current study. In addition, other anatomical boundaries present in the manual method protocol appear to match well with those of the automated ones. These areas include the superiolateral white matter bands of the hippocampus, the fornix and more anterior, the fimbria and the alveus. Also, both the manual and the automated segmentation method include at least part of the subiculum. Thus, differences anatomical boundaries between segmentation protocols could be expected to result in volume changes of around 5%, with the automated methods generating larger volumes than the manual one. This is clearly not what is seen, as the automated methods generate hippocampus volumes that are approximately 28% (FSL) and 55% (Freesurfer) larger than the manual ones. This additional overestimation could be the consequence of using a standard brain template derived from mature adult brains compared to a pediatric population. Future studies should determine whether using a common space based on pediatric brains, which would be more representative of this population’s neuroanatomy, could potentially improve the accuracy of automated segmentation techniques. Another possible cause for this additional discrepancy can be seen in Figure 6, which illustrates for selected subjects that both automated methods suffer from inclusion of ventricle space, neighbouring gray matter structures, and white matter in their segmentations. There are probably multiple reasons for the inclusion of these structures and areas not part of the target structure. Signal intensity might vary depending on scan quality and motion artefacts, which may lead to a less precise differentiation and classification of structures by automated techniques. This might be especially significant in children, who are more likely to move during scan acquisition. Although, we performed a visual quality control to remove scans with apparent motion artefacts, it cannot be excluded that motion affected the quality of the results from the two automatic segmentation protocols. Further, even in scans of high quality, superior and lateral boundaries of the amygdala with the basal ganglia, inferior boundaries with the hippocampus and lateral-inferior boundaries with the entorhinal cortex

can be difficult to define based on signal intensity, and are highly variable across subjects due to anatomical heterogeneity. Consequently, manual segmentation protocols often rely on the visualization of the area by a trained anatomist, recognition of the various structures in the field of view, and an expert decision as to where exactly the boundary to surrounding structures is located for that particular subject. This is a procedure that is time and labour intensive but favours anatomical precision and validity. Automated methods, in comparison, can't rely on an expert rater's decision in ambiguous circumstances, and have to employ probabilities and intensity distributions instead. Future studies investigating differences in the 3-D shape of the hippocampus and amygdala segmented manually or with automated techniques could allow a better understanding of the discrepancy in volume observed when comparing manual and automatic segmentation.

Volume overestimation does not necessarily imply a lack of validity of automatic segmentation as long as it is done in a consistent manner. Thus, to assess consistency in volume estimation, we computed pearson correlations between volumes derived from automatic methods and manual segmentation. The guiding idea was that a consistent overestimation of volumes would not weaken correlations between segmentation techniques and could thus still support the validity of automated techniques relative to manual segmentation. Associations between FreeSurfer and manual segmentation were satisfactory for the hippocampus volumes and ranged between r=0.69 to r=0.77. These correlations are consistent with what as previously been reported in the literature (Cherbuin et al., 2009; Doring et al., 2011; Morey et al., 2009; Pardoe et al., 2009; Pipitone et al., 2014; Sánchez-Benavides et al., 2010), which usually supports correlation coefficients surrounding r=0.75. Correlations between amygdala volumes derived FreeSurfer and manual segmentation were weaker than for the hippocampus and ranged between r=0.61 and r=0.67. These estimates are consistent with what has been found by Morey et al. (2009). However, few studies have looked at the accuracy of FreeSurfer to estimate amygdala volume, thus it is difficult to compare our results with previous findings. Past studies comparing manually segmented hippocampus volumes to volumes obtained with FSL-FIRST reported Pearson correlation coefficients varied between r=0.47(Pardoe et al., 2009) to r=0.67(Nugent et al., 2013). The results we obtained performing similar analyses highlight correlations closer to lower estimates that have been reported in the past (r=0.51 to r=059). While neuroanatomical characteristics of the studied pediatric population could have contributed to lower correlations found in this study, similar correlations between FSL-FIRST and manual segmentation have been highlighted in past studies performed on adults/mature brains (Pardoe et al., 2009; Doring et al., 2011). For amygdala volumes derived using FSL-FIRST, correlations with manual segmentation and FSL-FIRST can be considered weak (r=0.31 to r=035). The poor correlation between FSL-FIRST and manual segmentation for assessment of the amygdala volume has also been reported previously (Morey et al., 2009). The assessment of reproducibility of measurements with ICC suggests a weak agreement between manual segmentation and automated methods. The only comparison that reached or exceeded a coefficient of 0.70, a threshold previously defined as the minimum to define reliability between measures (Nunnally et al., 1967; Terwee et al., 2007), was the left hippocampus volume measured with FreeSurfer and manual segmentation. All other volumes, from FreeSurfer or FSL-FIRST, failed to reach this minimum standard to support

adequate agreement with manual segmentation. Two key observations could be derived from Pearson correlations and ICCs analyses. First, the agreement between manual and automated segmentation tended to be stronger for hippocampus than amygdala volumes. This amygdala-hippocampus discrepancy was also observed in previous articles studying the validity of automated segmentation in medial temporal lobe structures (Morey et al., 2009). Poor associations found with amygdala volumes are possibly the consequence of the neuroanatomical complexity of this structure. In addition to poor agreement with manual segmentation, the amygdala volume was also shown to have a low scan-rescan reliability when estimated with automatic techniques (Morey et al., 2010), most likely due to a high susceptibility to small variations in image intensity. This suggests that the amygdala volume is particularly difficult to assess reliably and vulnerable to errors when estimated with automated methods. The second observation that was noted both in between-method correlations and ICC analyses was that the association between automatic and manual segmentation volumes was stronger with FreeSurfer than FSL-FIRST. This was true for both the amygdala and hippocampus volume. Indeed, FreeSurfer consistently yielded larger correlations and ICC coefficients with manual segmentation than FSL-FIRST. It seems that the advantage of FreeSurfer over FSL-FIRST is not specific to our population as it was outlined in previous articles comparing results from both segmentation techniques as well (Doring et al., 2011; Morey et al., 2009; Pardoe et al., 2009). Using native manual segmentation volumes computed either by resampling labels to the native space or by dividing volumes of labels in the standard space by scale factors associated with the linear transformation did not significantly alter results of these analyses and did not change our findings (see Table 4).

An approach commonly used to establish the validity of automated techniques is to define their accuracy in distinguishing individuals from different clinical groups (eg. Alzheimer Disease versus Normal aging patients). A limitation associated with this study lies in the absence subgroups in the studied population. However, to arrive at an assessment of consistency for each method independently, we used between hemisphere correlations to demonstrate the differences between methods. If manual and automated segmentations were interchangeable, it would be expected that the associations between left and right volumes would be similar regardless of the difference in structural definition associated with the segmentation protocol. Further, it can be expected that, within the individual subject, left versus right hemispheric volumes are moderately to strongly associated (Allen et al., 2002). Both manual segmentation and FreeSurfer seemed to support this last statement, with results showing strong correlations between left versus right hemisphere for both the amygdala and the hippocampus. Correlations between bilateral amygdala and hippocampus volumes were significantly weaker when estimated with FSL-FIRST. Thus, within-method correlations suggest that the FSL-FIRST method might be prone to inconsistencies in segmentation within the same subject. The scope of this study was to investigate two key structures of the medial temporal lobe, the hippocampus and amygdala. While results highlighted in this article are likely to extend to adjacent structures in the medial temporal lobe, and perhaps to the rest of the cortex, our findings remain specific to those two key structures. Future studies investigating the agreement between manual and automated segmentation using a more

global approach and looking as spatial relationships between segmented structures would provide important additional information.

When looking at associations between manual segmentation and both automated techniques and the overall fit to the regression line, a considerable number of outliers could be visually identified. In these outliers, a marked discrepancy between automatic and manual segmentation volume estimates is observed - contributing to a limited explanation of variance. It is thus possible that the automated segmentation tends to be particularly inaccurate for some subjects. The Bland-Altman diagrams seem to support that notion by indicating a trend for larger volume difference between manual and automated segmentation for individuals with smaller structure volumes. Variations in scan quality or even in anatomy could contribute to this variability in performance. The hippocampus shape and volume are known to be highly variable across normal subjects (Bouix et al., 2005; Lupien et al., 2007). Studies looking at hippocampal shape in pediatric populations highlighted variations in the hippocampal shape over the course of normal development (Gogtay et al., 2006; Lin et al., 2013). Additionally, Gogtay and colleagues (2006) reveal important between-subject heterogeneity in the development of the hippocampal structure during brain development. Automated techniques are likely to be less flexible and accurate when dealing with irregular shape. On the other hand, an expert in neuroanatomy and hippocampus segmentation should not be affected by variance in shape. Future studies should aim to investigate the impact associated with variations in the shape of neuroanatomical structures in the context of automatic segmentation validation. To illustrate cases were there is an important discrepancy between automated and manual segmentation, we selected subjects that deviated from the regression line and visually compared labels obtained with both techniques. In addition to corroborating the overestimation reported in previous analyses, these images show a tendency for automated methods to miss the borders of target structures and expand into adjacent areas, including ventricular space. For those subjects, the obtained volume is not anatomically valid and should not be used in subsequent analyses. This highlights the importance of quality control and, when needed, corrections of labels obtained automatically. This process is time and labour intensive and is rarely performed thoroughly. Both FreeSurfer and FSL-FIRST include documentation and guidelines on quality control. However, to reflect the way groups lacking the training and expertise in anatomy would use these tools, we did not apply any form of correction of the labels derived from automated methods in the current article. A careful and informed quality control and manual corrections of automatically obtained labels by a trained individual would likely lead to significantly improved associations between manual segmentation and automatic techniques, especially when used in special populations like the one used in the current study.

5. Conclusion

In this study we highlight differences in volumes of structures segmented manually or obtained with automatic techniques, in this case FreeSurfer and FSL-FIRST. We provide evidence that, in a pediatric population, volumes obtained with those techniques might not always be equivalent to volumes obtained when manually segmented by an anatomical expert. This is especially true for more complex structures, such as the amygdala. Our results also support a better consistency between manual segmentation and FreeSurfer than

FSL-FIRST. With these results, we hope to emphasize the importance of performing quality control on volumes obtained automatically. A validated and well-established quality control protocol could significantly improve the correspondance between automatic and manual segmentation volumes.

Acknowledgements:

This work was supported by the NIH R01 HD 50662 (EPD), R01 HD 51852 (CAS), P50 MH 096889 (EPD, CAS). DS doctoral training is funded by Fonds de recherche du Québec - Santé (FRQS) doctoral award. JCP is supported by an FRQS Chercheur National salary award.

REFERENCES

Adolphs R, Tranel D, Damasio H, Damasio A, 1994 Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature 372, 669–672. [PubMed: 7990957]

Akhondi-Asl A, Jafari-Khouzani K, Elisevich K, Soltanian-Zadeh H, 2011 Hippocampal volumetry for lateralization of temporal lobe epilepsy: automated versus manual methods. Neuroimage 54, S218–S226. [PubMed: 20353827]

Allen JS, Damasio H, Grabowski TJ, 2002 Normal neuroanatomical variation in the human brain: An MRI-volumetric study. American Journal of Physical Anthropology 118, 341–358. [PubMed: 12124914]

Arnold SE, Trojanowski JQ, 1996 Human fetal hippocampal development: I. Cytoarchitecture, myeloarchitecture, and neuronal morphologic features. Journal of Comparative Neurology 367, 274–292. [PubMed: 8708010]

Bland JM, Altman D, 1986 Statistical methods for assessing agreement between two methods of clinical measurement. The lancet 327, 307–310.

Bonne O, Brandes D, Gilboa A, Gomori JM, Shenton ME, Pitman RK, Shalev AY, 2001 Longitudinal MRI study of hippocampal volume in trauma survivors with PTSD. The American journal of psychiatry 158, 1248. [PubMed: 11481158]

Bouix S, Pruessner JC, Collins DL, Siddiqi K, 2005 Hippocampal shape analysis using medial surfaces. Neuroimage 25, 1077–1089. [PubMed: 15850726]

Burdock EI, Fleiss JL, Hardesty AS, 1963 A NEW VIEW OF INTER- OBSERVER AGREEMENT1. Personnel Psychology 16, 373–384.

Buss C, Davis EP, Shahbaba B, Pruessner JC, Head K, Sandman CA, 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, E1312–E1319.

Campbell S, MacQueen G, 2004 The role of the hippocampus in the pathophysiology of major depression. Journal of Psychiatry and Neuroscience 29, 417. [PubMed: 15644983]

Chakravarty MM, Rapoport JL, Giedd JN, Raznahan A, Shaw P, Collins DL, Lerch JP, Gogtay N, 2014 Striatal shape abnormalities as novel neurodevelopmental endophenotypes in schizophrenia: A longitudinal study. Human brain mapping.

Cherbuin N, Anstey KJ, Reglade-Meslin C, Sachdev PS, 2009 In vivo hippocampal measurement and memory: a comparison of manual tracing and automated segmentation in a large community-based sample. PLoS One 4, e5265.

Collins DL, Neelin P, Peters TM, Evans AC, 1994 Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. Journal of computer assisted tomography 18, 192–205. [PubMed: 8126267]

Davis EP, Sandman CA, Buss C, Wing DA, Head K, 2013 Fetal glucocorticoid exposure is associated with preadolescent brain development. Biological psychiatry 74, 647–655. [PubMed: 23611262]

Davis M, Whalen PJ, 2001 The amygdala: vigilance and emotion. Molecular psychiatry 6, 13–34. [PubMed: 11244481]

Donegan NH, Sanislow CA, Blumberg HP, Fulbright RK, Lacadie C, Skudlarski P, Gore JC, Olson IR, McGlashan TH, Wexler BE, 2003 Amygdala hyperreactivity in borderline personality disorder: implications for emotional dysregulation. Biological psychiatry 54, 1284–1293. [PubMed: 14643096]

Doring TM, Kubo TT, Cruz LCH, Juruena MF, Fainberg J, Domingues RC, Gasparetto EL, 2011 Evaluation of hippocampal volume based on MR imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. Journal of Magnetic Resonance Imaging 33, 565–572. [PubMed: 21563239]

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, 2002 Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355. [PubMed: 11832223]

Fox N, Warrington E, Freeborough P, Hartikainen P, Kennedy A, Stevens J, Rossor MN, 1996 Presymptomatic hippocampal atrophy in Alzheimer's disease A longitudinal MRI study. Brain 119, 2001–2007. [PubMed: 9010004]

Giedd JN, Raznahan A, Alexander-Bloch A, Schmitt E, Gogtay N, Rapoport JL, 2015 Child psychiatry branch of the national institute of mental health longitudinal structural magnetic resonance imaging study of human brain development. Neuropsychopharmacology 40, 43–49. [PubMed: 25195638]

Gogtay N, Nugent TF, Herman DH, Ordonez A, Greenstein D, Hayashi KM, Clasen L, Toga AW, Giedd JN, Rapoport JL, 2006 Dynamic mapping of normal human hippocampal development. Hippocampus 16, 664–672. [PubMed: 16826559]

Herpertz SC, Dietrich TM, Wenning B, Krings T, Erberich SG, Willmes K, Thron A, Sass H, 2001 Evidence of abnormal amygdala functioning in borderline personality disorder: a functional MRI study. Biological psychiatry 50, 292–298. [PubMed: 11522264]

Jack C, Bentley M, Twomey C, Zinsmeister A, 1990 MR imaging-based volume measurements of the hippocampal formation and anterior temporal lobe: validation studies. Radiology 176, 205–209. [PubMed: 2353093]

Jack CR Jr, Theodore WH, Cook M, McCarthy G, 1995 MRI-based hippocampal volumetrics: Data acquisition, normal ranges, and optimal protocol. Magnetic resonance imaging 13, 1057–1064. [PubMed: 8750317]

Janssen I, Krabbendam L, Bak M, Hanssen M, Vollebergh W, Graaf R.d., Os J.v., 2004 Childhood abuse as a risk factor for psychotic experiences. Acta Psychiatrica Scandinavica 109, 38–45. [PubMed: 14674957]

Johnson JG, Cohen P, Brown J, Smailes EM, Bernstein DP, 1999 Childhood maltreatment increases risk for personality disorders during early adulthood. Archives of general psychiatry 56, 600–606. [PubMed: 10401504]

Krouwer JS, 2008 Why Bland-Altman plots should use X, not (Y+ X)/2 when X is a reference method. Statistics in medicine 27, 778–780. [PubMed: 17907247]

Lin M, Fwu PT, Buss C, Davis EP, Head K, Muftuler LT, Sandman CA, Su M-Y, 2013 Developmental changes in hippocampal shape among preadolescent children. International Journal of Developmental Neuroscience 31, 473–481. [PubMed: 23773912]

Lupien S, Evans A, Lord C, Miles J, Pruessner M, Pike B, Pruessner J, 2007 Hippocampal volume is as variable in young as in older adults: implications for the notion of hippocampal atrophy in humans. Neuroimage 34, 479–485. [PubMed: 17123834]

MacMillan HL, Fleming JE, Streiner DL, Lin E, Boyle MH, Jamieson E, Duku EK, Walsh CA, Wong MY-Y, Beardslee WR, 2001 Childhood abuse and lifetime psychopathology in a community sample. Childhood 158.

Matsuoka Y, Mori E, Inagaki M, Kozaki Y, Nakano T, Wenner M, Uchitomi Y, 2003 [Manual tracing guideline for volumetry of hippocampus and amygdala with high-resolution MRI]. No To Shinkei 55, 690–697. [PubMed: 13677303]

Meng X-L, Rosenthal R, Rubin DB, 1992 Comparing correlated correlation coefficients. Psychological bulletin 111, 172.

Morey RA, Petty CM, Xu Y, Pannu Hayes J, Wagner II HR, Lewis DV, LaBar KS, Styner M, McCarthy G, 2009 A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45, 855–866. [PubMed: 19162198]

Morey RA, Selgrade ES, Wagner HR, Huettel SA, Wang L, McCarthy G, 2010 Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Human brain mapping 31, 1751–1762. [PubMed: 20162602]

Nugent AC, Luckenbaugh DA, Wood SE, Bogers W, Zarate CA, Drevets WC, 2013 Automated subcortical segmentation using FIRST: Test-retest reliability, interscanner reliability, and comparison to manual segmentation. Human brain mapping 34, 2313–2329. [PubMed: 22815187]

Nunnally JC, Bernstein IH, Berge J.M.t., 1967 Psychometric theory. McGraw-Hill New York.

Pardoe HR, Pell GS, Abbott DF, Jackson GD, 2009 Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? Epilepsia 50, 2586–2592. [PubMed: 19682030]

Patenaude B, 2007 Bayesian Statistical Models of Shape and Appearance for Subcortical Brain Segmentation. University of Oxford.

Patenaude B, Smith SM, Kennedy DN, Jenkinson M, 2011 A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56, 907–922. [PubMed: 21352927]

Paus T, Keshavan M, Giedd JN, 2008 Why do many psychiatric disorders emerge during adolescence? Nature Reviews Neuroscience 9, 947–957. [PubMed: 19002191]

Pipitone J, Park MTM, Winterburn J, Lett TA, Lerch JP, Pruessner JC, Lepage M, Voineskos AN, Chakravarty MM, Initiative, A.s.D.N., 2014 Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. Neuroimage 101, 494–512. [PubMed: 24784800]

Pruessner J, Li L, Serles W, Pruessner M, Collins D, Kabani N, Lupien S, Evans A, 2000 Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cerebral Cortex 10, 433–442. [PubMed: 10769253]

Pruessner JC, Köhler S, Crane J, Pruessner M, Lord C, Byrne A, Kabani N, Collins DL, Evans AC, 2002 Volumetry of temporopolar, perirhinal, entorhinal and parahippocampal cortex from high-resolution MR images: considering the variability of the collateral sulcus. Cerebral Cortex 12, 1342–1353. [PubMed: 12427684]

Pynoos RS, Steinberg AM, Piacentini JC, 1999 A developmental psychopathology model of childhood traumatic stress and intersection with anxiety disorders. Biological psychiatry 46, 1542–1554. [PubMed: 10599482]

Raghunathan TE, Rosenthal R, Rubin DB, 1996 Comparing correlated but nonoverlapping correlations. Psychological Methods 1, 178.

Rauch SL, Whalen PJ, Shin LM, McInerney SC, Macklin ML, Lasko NB, Orr SP, Pitman RK, 2000 Exaggerated amygdala response to masked facial stimuli in posttraumatic stress disorder: a functional MRI study. Biological psychiatry 47, 769–776. [PubMed: 10812035]

Raznahan A, Shaw PW, Lerch JP, Clasen LS, Greenstein D, Berman R, Pipitone J, Chakravarty MM, Giedd JN, 2014 Longitudinal four-dimensional mapping of subcortical anatomy in human development. Proceedings of the National Academy of Sciences 111, 1592–1597.

Rodionov R, Chupin M, Williams E, Hammers A, Kesavadas C, Lemieux L, 2009 Evaluation of atlas-based segmentation of hippocampi in healthy humans. Magnetic resonance imaging 27, 1104–1109. [PubMed: 19261422]

Sánchez-Benavides G, Gomez-Anson B, Sainz A, Vives Y, Delfino M, Pena-Casanova J, 2010 Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. Psychiatry Research: Neuroimaging 181, 219–225. [PubMed: 20153146]

Shen L, Saykin AJ, Kim S, Firpi HA, West JD, Risacher SL, McDonald BC, McHugh TL, Wishart HA, Flashman LA, 2010 Comparison of manual and automated determination of hippocampal volumes in MCI and early AD. Brain imaging and behavior 4, 86–95. [PubMed: 20454594]

Shrout PE, Fleiss JL, 1979 Intraclass correlations: uses in assessing rater reliability. Psychological bulletin 86, 420. [PubMed: 18839484]

Sled JG, Zijdenbos AP, Evans AC, 1998 A nonparametric method for automatic correction of intensity nonuniformity in MRI data. Medical Imaging, IEEE Transactions on 17, 87–97.

Springer KW, Sheridan J, Kuo D, Carnes M, 2007 Long-term physical and mental health consequences of childhood physical abuse: Results from a large population-based sample of men and women. Child abuse & neglect 31, 517–530. [PubMed: 17532465]

Squire LR, 1992 Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychological review 99, 195. [PubMed: 1594723]

Steiger JH, 1980 Tests for comparing elements of a correlation matrix. Psychological bulletin 87, 245.

Stein MB, Goldin PR, Sareen J, Zorrilla LTE, Brown GG, 2002 Increased amygdala activation to angry and contemptuous faces in generalized social phobia. Archives of general psychiatry 59, 1027. [PubMed: 12418936]

Tae WS, Kim SS, Lee KU, Nam E-C, Kim KW, 2008 Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. Neuroradiology 50, 569–581. [PubMed: 18414838]

Teicher MH, Andersen SL, Polcari A, Anderson CM, Navalta CP, Kim DM, 2003 The neurobiological consequences of early stress and childhood maltreatment. Neuroscience & Biobehavioral Reviews 27, 33–44. [PubMed: 12732221]

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC, 2007 Quality criteria were proposed for measurement properties of health status questionnaires. Journal of clinical epidemiology 60, 34–42. [PubMed: 17161752]

Watson C, Andermann F, Gloor P, Jones-Gotman M, Peters T, Evans A, Olivier A, Melanson D, Leroux G, 1992 Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. Neurology 42, 1743–1743. [PubMed: 1513464]

Woon FL, Hedges DW, 2008 Hippocampal and amygdala volumes in children and adults with childhood maltreatment- related posttraumatic stress disorder: A meta- analysis. Hippocampus 18, 729–736. [PubMed: 18446827]

Highlights:

- In a pediatric population, we compare hippocampus and amygdala volumes obtained with FSL-FIRST and FreeSurfer to volumes obtained with manual segmentation
- We examine discrepancies, associations, and biases between automatic and manual segmentation volumes
- In the studied pediatric population, the agreement between manual segmentation, FreeSurfer and FSL is questionable
- Associations between volumes derived from manual segmentation and FreeSurfer were stronger than with volumes derived from FSL-FIRST
- Associations between volumes derived from manual segmentation and automatic techniques were stronger for hippocampus than amygdala volumes

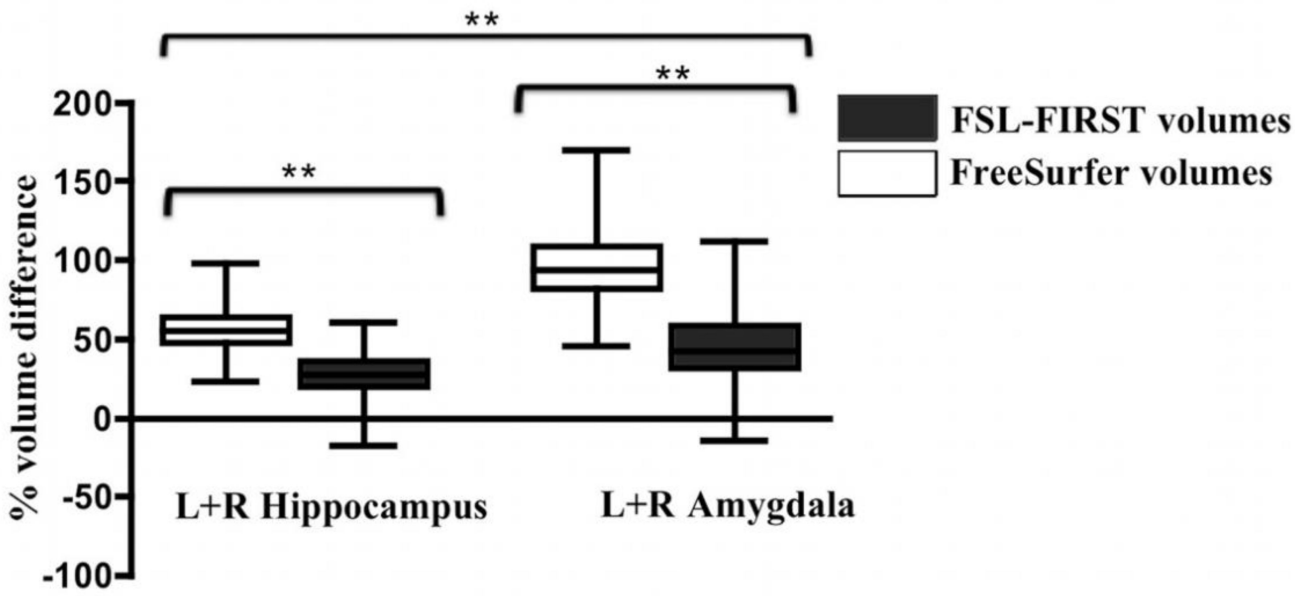

Fig. 1.

Percentage of volume difference between automatic protocols and manual segmentation for the combined left and right hippocampus and amygdala volumes. Two asterix indicate a significant difference (at the $p<0.0001$ level). Percent volume differences are significantly larger for volumes estimated with FreeSurfer than FSL-FIRST, for both the amygdala and the hippocampus. Further, the amygdala leads to significantly larger percent volume differences than the hippocampus, for FreeSurfer and FSL-FIRST.

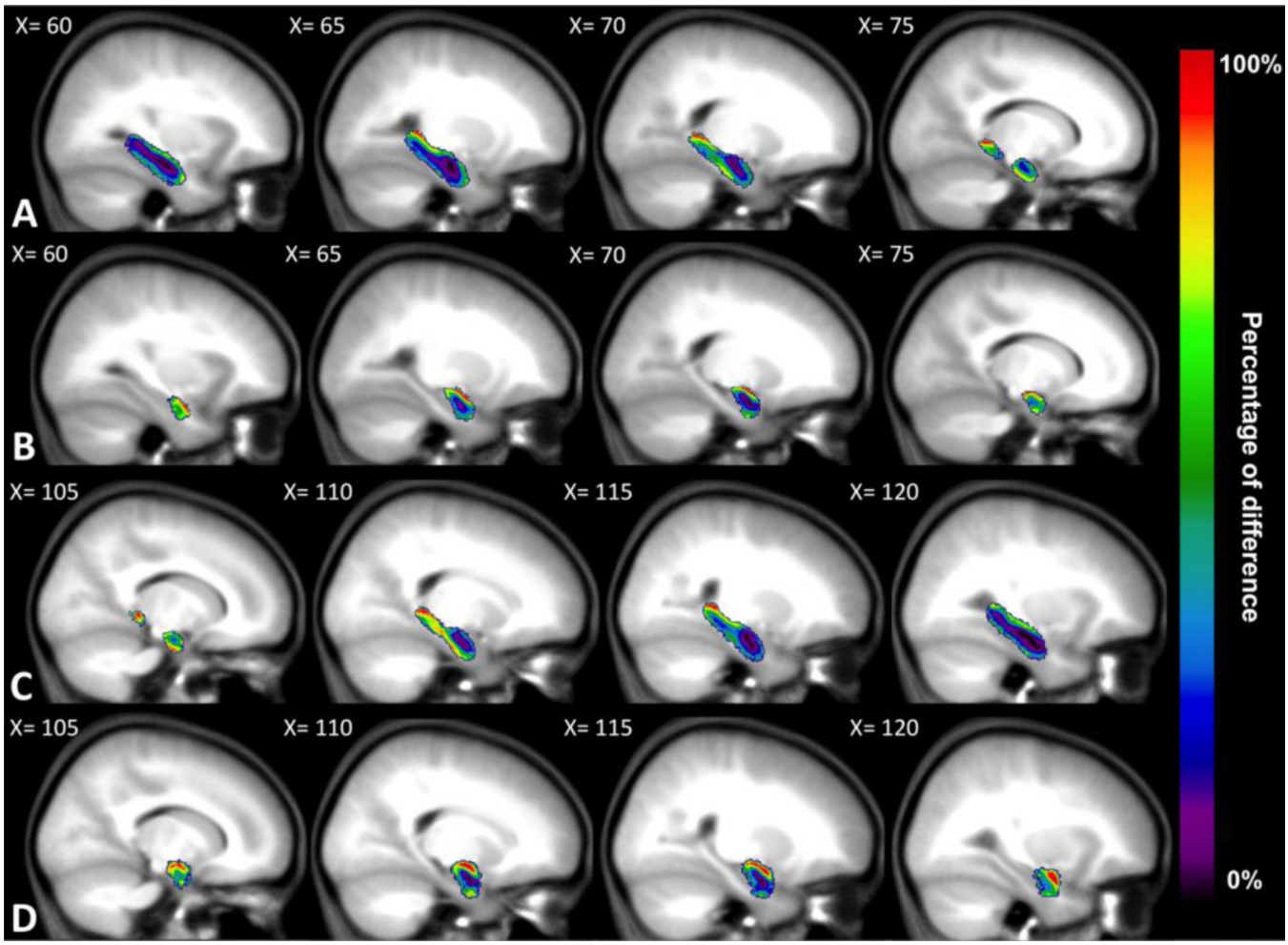

Fig. 2.

Satistical maps representing, for each voxel, the average percentage of difference between manual segmentation and FSL-FIRST volumes for the A-left hippocampus, B- left amygdala, C-right hippocampus, D-right amygdala. The maps are displayed on the average standardized brain of all subjects.

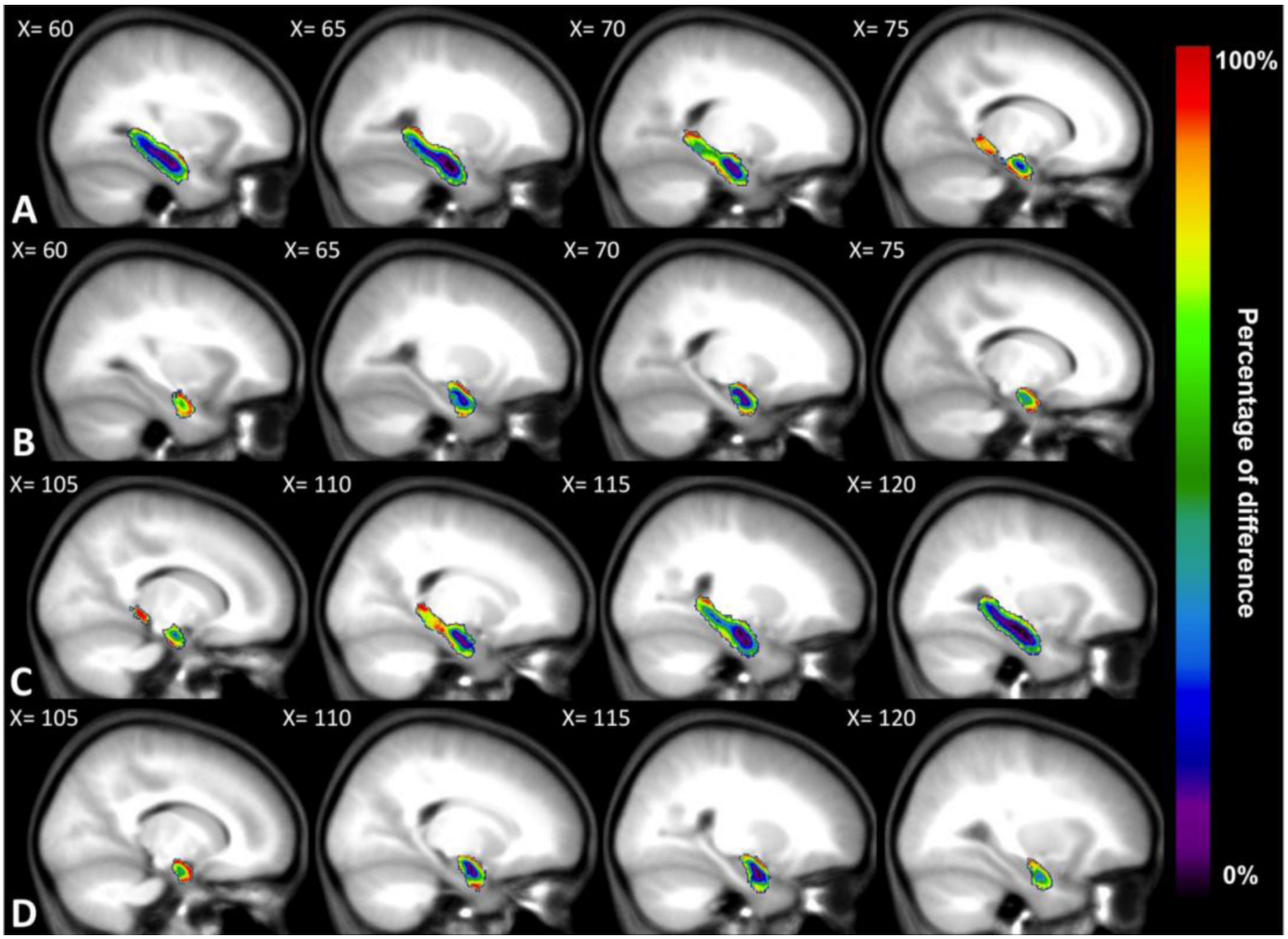

Fig. 3.

Satistical maps representing, for each voxel, the average percentage of difference between manual segmentation and Freesurfer volumes for the various structures. A-left hippocampus, B- left amygdala, C-right hippocampus, D-right amygdala. The maps are displayed on the average standardized brain of all subjects.

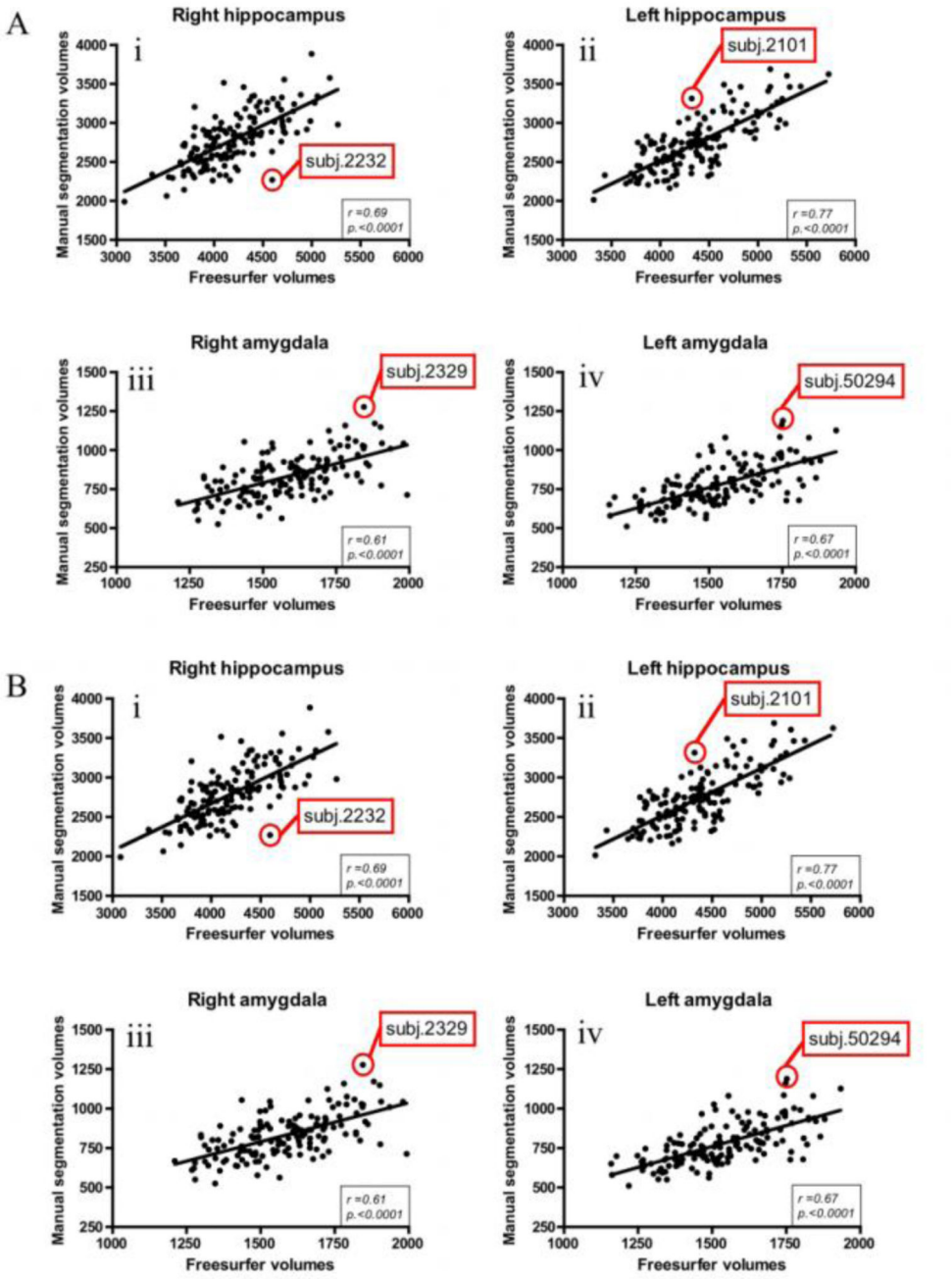

Fig. 4.

Pearson correlations between volumes obtained with manual segmentation and with FreeSurfer (A) and FSL-FIRST (B). Plots are presented separately for i-right hippocampus ii- left hippocampus, iii-right amygdala, iv-left amygdala. *r* - pearson correlation coefficient. Outliers, defined using the magnitude of the residuals, are circled in red and identified in a red rectangle.

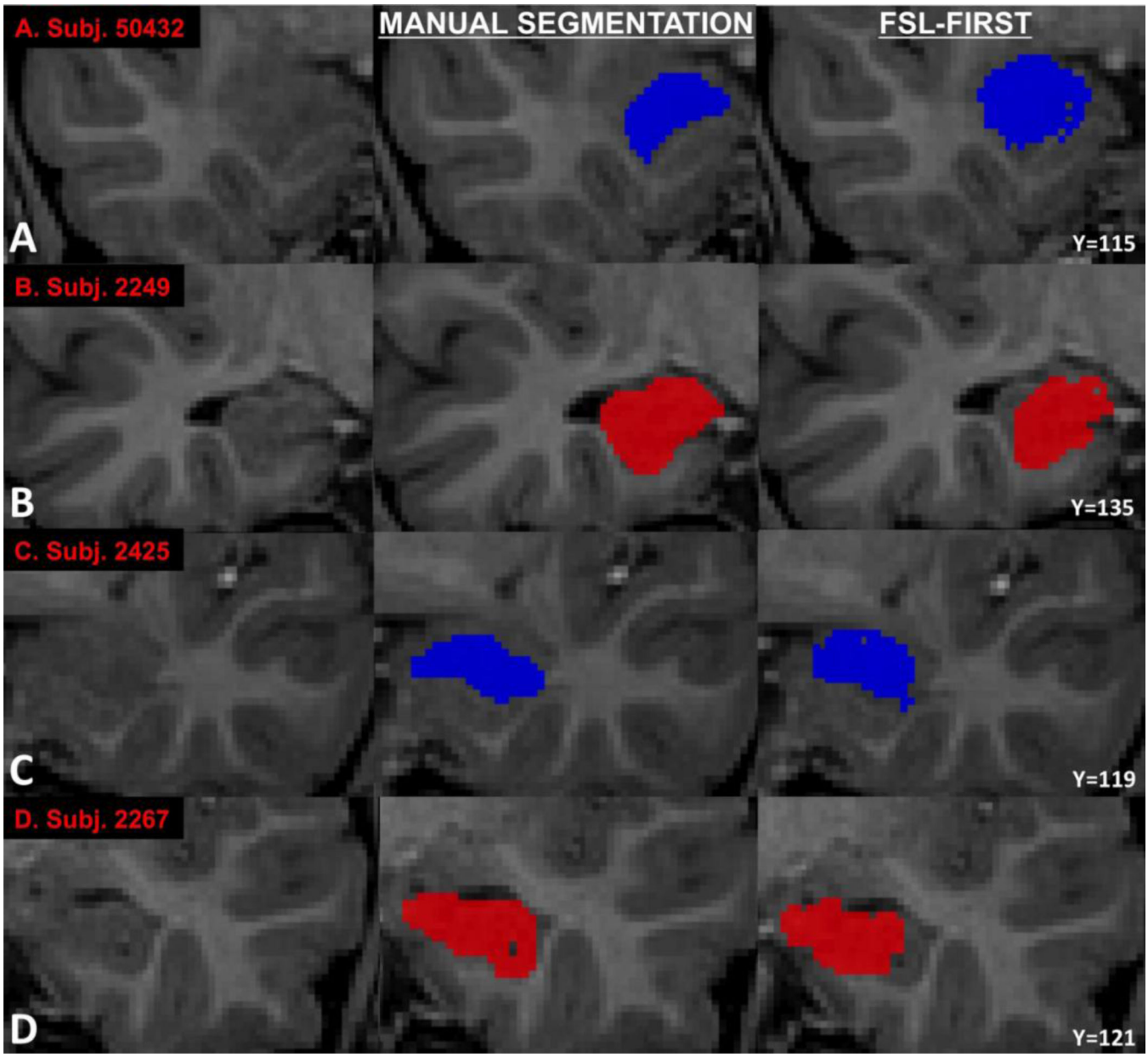

Fig. 5.

Visual comparison of amygdala (blue) and hippocampus (red) volume estimation in a single subject using manual segmentation and FSL-FIRST. These subjects were selected on the basis of linear regression analyses, due to a poor correspondance between manually segmented and FSL-FIRST derived volumes. A-left amygdala, B- left hippocampus, C-right amygdala, D-right hippocampus.

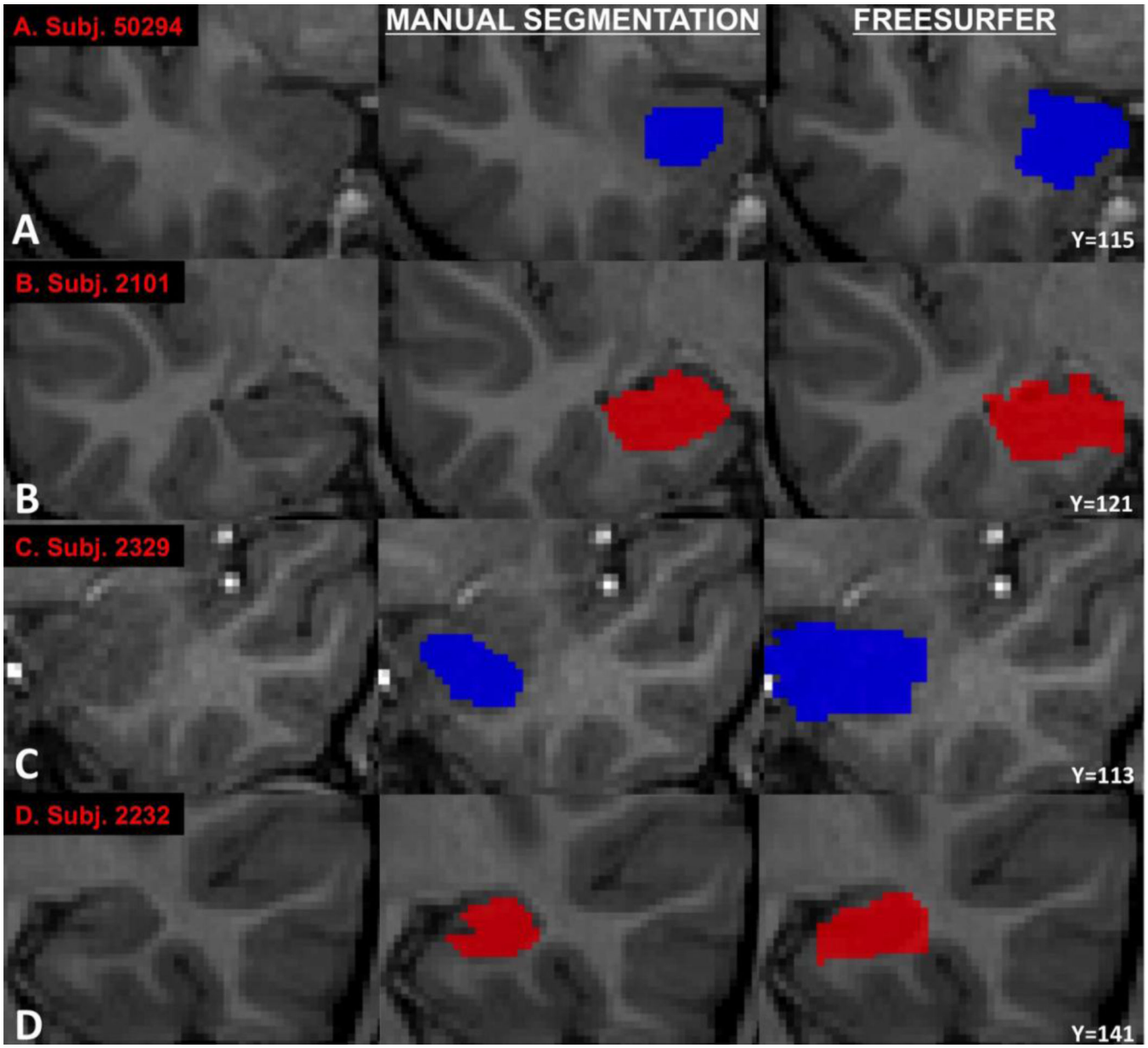

Fig. 6.

Visual comparison of amygdala (blue) and hippocampus (red) volume estimation in a single subject using manual segmentation and FreeSurfer. These subjects were selected on the basis of linear regression analyses, due to a poor correspondence between manually segmented and FreeSurfer derived volumes. A-left amygdala, B- left hippocampus, C-right amygdala, D-right hippocampus.

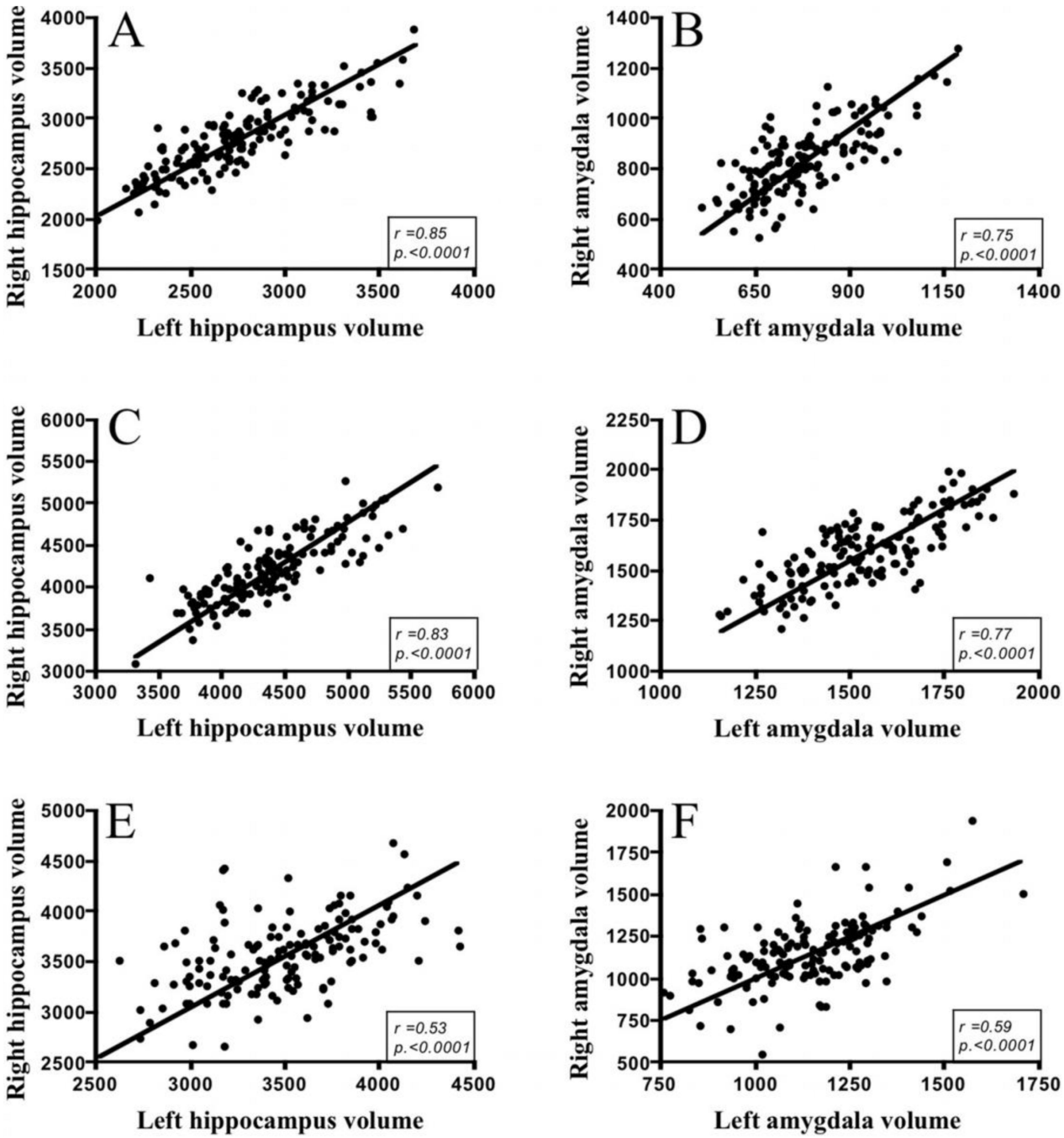

Fig. 7.

Within-method correlations of left versus right structure volumes. A- Correlation between left and right hippocampus volumes segmented manually B- Correlation between left and right amygdala volumes segmented manually C- Correlation between left and right hippocampus volumes segmented automatically with FreeSurfer D- Correlation between left and right amygdala volumes segmented automatically with FreeSurfer E- Correlation between left and right hippocampus volumes segmented automatically with FSL-FIRST D-

Correlation between left and right amygdala volumes segmented automatically with FSL-FIRST. *r* - pearson correlation coefficient.

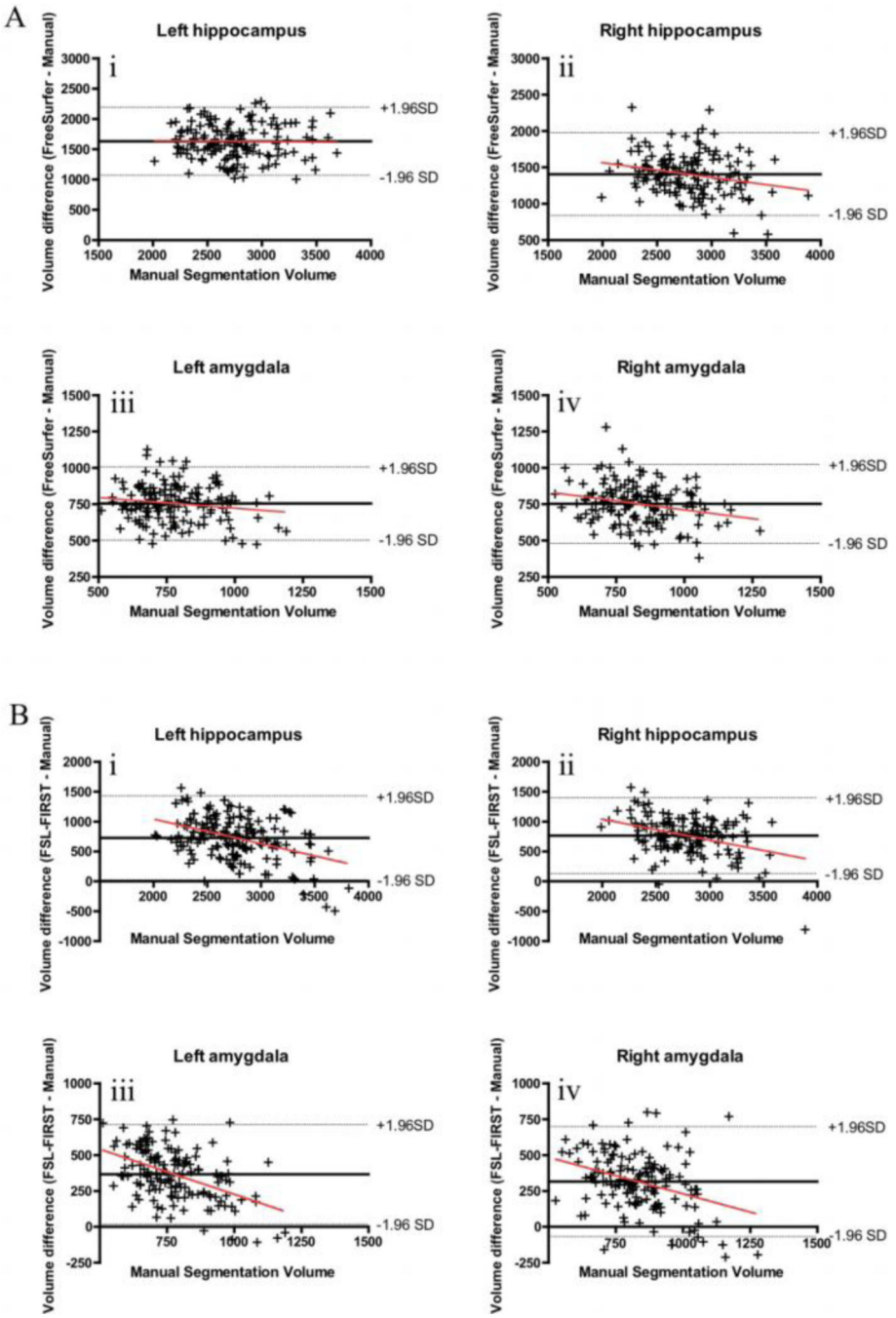

Fig. 8.

Bland-Altman plots of volume difference estimation between manual segmentation and (A) FreeSurfer or (B) FSL-FIRST. Plots are presented separately for i-left hippocampus ii- right hippocampus, iii-left amygdala, iv-right amygdala. A red regression line was integrated to each plot to illustrate potential biases in volume estimation.

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Demographic information

	Mean (SD)
N	**147**
Age	**8,47 (1,37)**
Gender (M/F)	**82/65**
Handedness (R/L)	**130/17**

Subject demographics. M — male. F — female. R— right handed. R— left handed. SD— standard deviation.

Table 2

Comparison of volumes between methods

	Manual	FSL-FIRST		FreeSurfer	
	Volume (SD)	Volume (SD)	% volume diff. (SD)	Volume (SD)	% volume diff. (SD)
L-hippocampus	**2746,29 (347,73)**	**3475,44 (378,37)**	**27.61 (14.49)**	**4378,05 (445,69)**	**60.38 (13.04)**
R-hippocampus	**2786,92 (337,88)**	**3553,06 (372,55)**	**28.39 (13.07)**	**4194,63 (390,10)**	**51.53 (13.17)**
L-amygdala	**777,27 (134,99)**	**1144,02 (163,19)**	**50.32 (27.65)**	**1532,65 (171,49)**	**100.29 (24.56)**
R-amygdala	**832,92 (137,71)**	**1148,97 (194,78)**	**40.29 (26.09)**	**1586,05 (170,55)**	**93.56 (25.78)**

Description of mean volumes derived from each technique as well as mean percentage of volume difference (% volume diff.) obtained between FreeSurfer/FSL-FIRST and manual segmentation. L — left. R — right. SD— standard deviation.

Table 3

Comparison of inter-hemispheric volumes correlations derived from each method

	Left-right hippocampus	Left-right amygdala
Manual segmentation	**0.85**	**0.75**
FreeSurfer	**0.83**	**0.77**
FSL-FIRST	**0.53** **	**0.59** **

Pearson correlations of left against right hemispheric volumes obtained within a same segmentation method.

** indicates a significant difference (at the $p<0.0001$ level) in the magnitude of the correlation, as compared with the correlation coefficients obtained with manual segmentation, as defined with the Fisher r-to-Z transform (ZPF) statistical test.

Table 4

Comparison of results obtained using native volumes derived from label resampling and scaling factor correction.

	Native manual segmentation volumes based on resampling of labels (as presented in the article)		Native manual segmentation volumes based on calculation of scaling factors	
	FreeSurfer	**FSL-FIRST**	**FreeSurfer**	**FSL-FIRST**
% volume diff. (SD)				
L-hippocampus	**60.38 (13.04)**	**27.61(14.49)**	**60.16 (12.90)**	**27.43 (14.28)**
R-hippocampus	**51.53 (13.17)**	**28.39 (13.07)**	**51.41 (13.09)**	**28.28 (12.89)**
L-amygdala	**100.29 (24.56)**	**50.32 (27.65)**	**99.53 (24.27)**	**49.76 (27.55)**
R-amygdala	**93.56 (25.78)**	**40.29 (26.09)**	**93.11 (24.56)**	**40.03 (25.83)**
PCC with manual seg.				
L-hippocampus	**0.77**	**0.51**	**0.77**	**0.52**
R-hippocampus	**0.69**	**0.59**	**0.70**	**0.60**
L-amygdala	**0.67**	**0.31**	**0.66**	**0.30**
R-amygdala	**0.61**	**0.35**	**0.62**	**0.35**
ICC with manual seg.				
L-hippocampus	**0.74**	**0.51**	**0.74**	**0.52**
R-hippocampus	**0.68**	**0.59**	**0.69**	**0.60**
L-amygdala	**0.65**	**0.30**	**0.64**	**0.28**
R-amygdala	**0.60**	**0.33**	**0.61**	**0.33**

Percentage of volume difference (% volume diff.), Pearson correlation coefficients (PCC) and intraclass correlation coefficients (ICC) computed between manual segmentation volumes and automatic protocols. Results are presented with native manual segmentation volumes obtained by resampling labels in the native space using an inversion of the linear transformation (left column) and with manual volumes obtained by dividing volumes segmented in the standard space by scale factors associated with the linear transformation in the x,y,z directions (right column). This table shows that both methods of estimating manual segmentation volumes in the native space lead to highly similar results. L — left. R — right. SD— standard deviation.