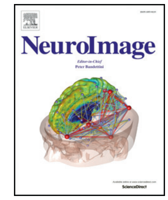
NeuroImage xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



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, CA, USA
[d] Charité, Berlin, Germany
[e] University of Denver, CO, USA
[f] Biomedical Engineering Department, McGill University, Montreal, QC, Canada

A R T I C L E I N F O

Article history:
Received 21 August 2015
Accepted 15 January 2016
Available online xxxx

Keywords:
Segmentation techniques
Pediatric population
Hippocampus
Amygdala
FSL-FIRST
FreeSurfer

A B S T R A C T

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 $r = 0.35$, $p < 0.001$; 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.

1. Introduction

Childhood is a period of great relevance in the development of risk factors for various neuropsychiatric conditions (Paus et al., 2008).

* Corresponding author at: McGill Centre for Studies in Aging, 6825 Boulevard LaSalle, Montreal, QC H4H1R3, Canada.

http://dx.doi.org/10.1016/j.neuroimage.2016.01.038
1053-8119/© 2016 Published by Elsevier Inc.

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 pediatric 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 labeled 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

Please cite this article as: Schoemaker, D., et al., Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manu..., NeuroImage (2016), http://dx.doi.org/10.1016/j.neuroimage.2016.01.038