**Manuscript**

**A comparative study of segmentation techniques for the quantification of brain subcortical volume**

**Authors** Theophilus N. Akudjedu1\*, Leila Nabulsi1, Migle Makelyte1, 2, Cathy Scanlon1, Sarah Hehir1, Helen Casey1, Srinath Ambati1, Joanne Kenney1, Stefani O’Donoghue1, Emma McDermott1, Liam Kilmartin2, Peter Dockery1, Colm McDonald1, Brian Hallahan1, Dara M. Cannon1.

**Affiliations** 1Centre for Neuroimaging & Cognitive Genomics (NICOG), Clinical Neuroimaging Laboratory, NCBES Galway Neuroscience Centre, Psychiatry & Anatomy, School of Medicine,College of Medicine Nursing and Health Sciences, National University of Ireland Galway, H91TK33 Galway, Ireland.

2College of Engineering and Informatics, National University of Ireland Galway, H91TK33 Galway, Ireland.

**Keywords** Segmentation techniques, Stereology, subcortical structures, FreeSurfer, FSL-FIRST, volBrain

**Acknowledgements** TNA’s doctoral training is funded by the College of Medicine, Nursing and Health Sciences Postgraduate Scholarship Scheme, NUI Galway (2016-2020). We would also like to thank all of the participants and their families for their involvement in the Research Programme of the Clinical Neuroimaging laboratory, NUI Galway.

**Corresponding Author**

Theophilus N. Akudjedu

Centre for Neuroimaging & Cognitive Genomics (NICOG), Clinical Neuroimaging Laboratory, NCBES Galway Neuroscience Centre, Psychiatry & Anatomy, School of Medicine,College of Medicine Nursing and Health Sciences, National University of Ireland Galway, H91TK33 Galway, Ireland.

**Email** rasningo@gmail.com

**Abstract**

Manual tracing of magnetic resonance imaging (MRI) represents the gold standard for segmentation in clinical neuropsychiatric research studies, however automated approaches are increasingly used due to its time limitations. The accuracy of segmentation techniques for subcortical structures has not been systematically investigated in large samples.

We compared the accuracy of fully automated [(i) model-based: FSL-FIRST; (ii) patch-based: volBrain], semi–automated (FreeSurfer) and stereological (Measure®) segmentation techniques with manual tracing (ITK-SNAP) for delineating volumes of the caudate (*easy-to-segment*) and the hippocampus (*difficult-to-segment*).

High resolution 1.5 Tesla T1-weighted MR images were obtained from 177 patients with major psychiatric disorders and 104 healthy participants. The relative consistency (partial correlation), absolute agreement (intraclass correlation coefficient,ICC) andpotential technique bias (Bland-Altman plots)of each technique was compared with manual segmentation.

Each technique yielded high correlations (0.77-0.87, p<0.0001) and moderate ICC’s (0.28–0.49) relative to manual segmentation for the caudate. For the hippocampus, stereology yielded good consistency (0.52-0.55, p<0.0001) and ICC (0.47-0.49), whereas automated and semi-automated techniques yielded poor ICC (0.07-0.10) and moderate consistency (0.35-0.62, p<0.0001). Bias was least using stereology for segmentation of the hippocampus and using FreeSurfer for segmentation of the caudate.

In a typical neuropsychiatric MRI dataset, automated segmentation techniques provide good accuracy for an easy-to-segment structure such as the caudate, whereas for the hippocampus, a reasonable correlation with volume but poor absolute agreement was demonstrated. This indicates manual or stereological volume estimation should be considered for studies that require high levels of precision such as those with small sample size.

**Introduction**

In recent decades, significant advances in *in-vivo* brain imaging technologies have allowed for volumetric quantification of brain structures in large magnetic resonance imaging (MRI) datasets. This has enabled researchers to undertake large-scale world-wide studies examining brain anatomy and disease progression using morphometric properties. Several changes in cortical and subcortical morphometry are widely reported from clinical groups using segmentation techniques. Clinical neuromorphometric studies of cohorts including bipolar disorder (Quigley et al. 2015; Hibar et al. 2016; Sacchet et al. 2015; Mamah et al. 2016), first-episode psychosis (Scanlon et al. 2014; Cahn et al. 2002; Mamah et al. 2012), major depressive disorder (Schmaal et al. 2016; Sacchet et al. 2015; Renteria et al. 2017) and schizophrenia (van Erp et al. 2016; Okada et al. 2016; Mamah et al. 2016) have all employed various classes of segmentation processing techniques.

Manual segmentation techniques are currently considered the gold standard for volumetric quantification of regional brain structures (Morey et al. 2009; Rodionov et al. 2009; Doring et al. 2011; Sanchez-Benavides et al. 2010). However, such procedures are associated with a number of limitations including the requirement for substantial anatomical and methodological expertise by the tracer(s), with associated inter- and intra-rater reliability concerns, although the strict use of protocols combined with anatomical expertise and training can minimise these. In addition, manual tracing is associated with significant time requirements, potentially limiting the feasibility of this technique for the volumetric quantification of large datasets. Various automated techniques are now freely/commercially available to segment regional brain structures in order to overcome these limitations and have been employed in several large recent MRI collaborative studies (van Erp et al. 2016; Hibar et al. 2016; Franke et al. 2016).

Automated approaches for brain processing and segmentation have several potential advantages over manual segmentation. Advantages include a significant reduction in time and labour, excellent reproducibility and less training or anatomical knowledge for the end-user. Two automated, model-based approaches commonly used in the field currently are FreeSurfer (<http://surfer.nmr.mgh.harvard.edu>, Fischl et al. 2002) and "FIRST", which is provided as part of the FSL software library (<http://fsl.fmrib.ox.ac.uk>, Patenaude et al.2011). FreeSurfer automatically assigns a label to each voxel from anatomical images based on probabilistic estimations relying on Markov random fields (MRFs). FIRST uses a probabilistic framework with boundaries of brain structures based on signal intensity of T1-images and the expected shape of structures to be segmented. These model-based approaches assume that algorithms can reliably delineate anatomical regions irrespective of inter-individual differences or pathological changes in neuroanatomy and/or challenges of MR brain acquisition. In an attempt to address these concerns, fully automated multi-atlas label fusion approaches such as volBrain (<http://volbrain.upv.es>, Manjon and Coupe 2016) have also been implemented. FreeSurfer additionally allows the user to edit the quality of segmentation determined automatically, unlike FIRST and volBrain, and can thus be considered a “semi-automated” technique.

Furthermore, stereology, a long-standing reliable method for the estimation of volume of structures, has been employed with MR imaging and previously been demonstrated to deliver consistent findings when compared to manual tracing for both smooth (Ertekin et al. 2015) and complex (Sheline et al. 1999) structures.

A number of previous studies (Sheline et al. 1999; Pardoe et al. 2009; Morey et al. 2009; Doring et al. 2011; Keller et al. 2012; Ertekin et al. 2015; Schoemaker et al. 2016; Makowski et al. 2017) have assessed the performance of automated methods and/or stereology compared to manual segmentation in relatively small populations. However, to our knowledge, no previous study has compared in such a large cohort of individuals across this wide range of segmentation techniques (fully-automated, semi-automated and stereology). The aim of the present analysis is to examine the performance of fully-automated (model-based) (FSL-FIRST v.5.0.4), fully-automated (patch-based) (volBrain v.1.0), semi–automated (FreeSurfer v.5.1) and stereological (Measure®) segmentation techniques against manual segmentation (ITK-SNAP v.1.8) (<http://www.itksnap.org/>) for the quantification of the caudate andhippocampus in a large heterogeneous sample of adult MR images. The caudate and the hippocampus were chosen as two representative structures that provide a smooth (*easy-to-segment*) and complex (*difficult-to-segment*) (Grimm et al. 2015) boundary for testing, respectively. Additionally, volumetric abnormalities in these regions have been implicated in numerous neuropsychiatric disorders.

**Methods**

***Participants***

Structural MR images (n=281), were obtained from individuals aged 16 – 60 years of age as part of the Clinical Neuroimaging Laboratory Research Programme at the National University of Ireland, Galway (NUI Galway). Participants included 189 males (67.3%) and had a mean age of 36 years (± 11 SD). They consisted of individuals with major psychiatric disorders (n=177) and healthy participants (n=104). All participants provided written informed consent for the relevant studies and ethical approval was obtained from the NUI Galway, and Galway University Hospitals Research Ethics Committees.

***MRI Data Acquisition***

Clinical assessments and baseline protocols employed during recruitment have been described previously (Emsell et al. 2013; McFarland et al. 2013; Scanlon et al. 2014; Ahmed et al. 2015; Kenney et al. 2015). Briefly, MR data were acquired at University Hospital Galway (UHG) on a 1.5 Tesla Siemens Magnetom Symphony scanner (Erlangen, Germany) equipped with a 4-channel head coil. A standard localiser and T1-sagittal sequence was used to confirm participants’ radiological position and placement of the image field-of-view in alignment with the anterior-posterior commissure (AC-PC) line. The weighted magnetisation-prepared rapid acquisition of gradient echo (MPRAGE) sequence was employed to acquire volumetric T1-weighted images (160 slices) with the following parameters: repetition time (TR): 1140 ms, echo time (TE): 4.38 ms, inversion time (TI): 600 ms, flip angle 15; matrix size 256×256; an in-plane pixel size of 0.9 mm×0.9 mm and slice thickness of 0.9 mm.

***Image Pre-processing and Quality Analyses***

The volumetric T1 images were visually examined for quality and underwent correction for intensity non-uniformity as described previously (Emsell et al. 2013; Scanlon et al. 2014; Ahmed et al. 2015). Non-parametric non-uniform intensity normalisation (N3) was performed to correct for intensity non-uniformity (Sled et al. 1998).

***Volumetric Quantification using Manual Segmentation with ITK-SNAP***

The bilateral caudate nuclei and hippocampi were manually segmented by anatomically trained raters [JK, SA, SO and SH] using ITK-SNAP (v.1.8.0, Yushkevich et al. 2006) with inter-rater reliability [Caudate (Left: 0.97; Right: 0.96) and Hippocampus (Left: 0.89; Right: 0.94)] and intra-rater reliability scores greater than 0.85 for all raters.

ITK-SNAP displays the three orthogonal views of the MR image and a 3D view of the segmented structure. The pre-processed images underwent segmentation according to a strict anatomical protocol adapted from Looi et al. for the caudate (Looi et al. 2008) and from Velakoulis et al. and Strakowski et al. in the case of the hippocampus (Velakoulis et al. 2006; Strakowski et al. 1999).

The first slice included for the caudate was defined as that axial slice in which the grey matter (GM) of the caudate head was clearly distinguished from the surrounding frontal white matter (WM) anteriorly, the anterior commissure posteriorly, the internal capsule laterally, and a thin strip of WM medially. Tracing continued superiorly on every slice in which the GM of the caudate was bordered medially by the anterolateral portion of the lateral ventricle, anteriorly by the frontal WM and dorso-laterally by the internal capsule. The last slice was defined as the point at which the caudate body was no longer clearly visible laying medial to the lateral ventricle. Due care was taken to avoid inclusion of the nucleus accumbens.

The hippocampus was defined as consisting of the hippocampus proper (Ammon’s horn), the dentate gyrus, and a portion of the subiculum. In order to limit inclusion to hippocampal GM, the alveus, fimbria, fornix and entorhinal cortex were excluded. Tracing was performed in the three orthogonal views mainly in the sagittal and coronal views with the axial view providing guidance. The anterior boundary was demarcated by the alveus acting as a WM border over the hippocampal head (Strakowski et al. 1999). The posterior boundary was defined by referring to the crus with tracing ending at the point where the greatest continuation of the fornix was visible as grey switches to WM (Velakoulis et al. 2006). The medial boundary was indicated by the open end of the hippocampal and uncal fissures and the medial aspect of the ambiens gyrus. The uncus and the ambient gyrus lie between the hippocampal complex and the ambient cistern in initial slices (Velakoulis et al. 2006). The lateral boundary was formed by the inferior horn of the lateral ventricle and adjacent WM (Altshuler et al. 1998, 2000). The superior boundary was the alveus and the inferior boundary was the WM between the hippocampus and the underlying parahippocampal gyrus (Brambilla et al. 2003).

Following segmentation according to the standardised protocols, the segmented regions were magnified by 7pixels/mm3 simultaneously in all planes to examine accuracy of the segmentation surface. The segmented structures were further labelled (L/R) to maintain appropriate lateralisation according to their position. ITK-SNAP measured the volume of the structure by assigning each voxel a single label depending on whether the voxel was manually traced or not (Yushkevich et al. 2006). Thus, the number of voxels present within the traced area was used to compute the volume of the structure.

***Volumetric Quantification using Stereological Segmentation with Measure®***

Stereology is a set of methodologies that ensure rigorous analysis of size, shape, and other quantitative parameters such as length, surface area and volume of three dimensional objects based on observations in two dimensional sections (Garcia et al. 2007). Stereology has been employed as a novel alternative volumetric analysis technique for radiological anatomy. This technique is based on the Cavalieri principle and is robust, unbiased and efficient as it requires no assumptions about the structure (orientation or shape) under assessment (Gundersen et al. 1988).

Following this principle, the slices of the brain in which the caudate/hippocampus is present were examined. Careful observation by expert neuroanatomists [DC and CS] and a trained rater [EMcD] indicated that both structures appeared on approximately 60 slices of the images used for the study. Similarly, 10 randomly selected subjects were used to determine the intra-rater reliability [EMcD] for this study. The intra-rater reliability scores were 0.94 and 0.80 for the caudate and hippocampus respectively. The first slice was chosen randomly while the subsequent slices were selected at systematic intervals from the images. The grid size, number of slices, slice thickness and the structural boundaries for the caudate and the hippocampus were explicitly defined as per the manual segmentation technique [caudate (Looi et al. 2008) and hippocampus (Velakoulis et al. 2006 and Strakowski et al. 1999)]. A computer-generated point counting grid (stereological probe) was randomly superimposed on the image to automatically count and estimate the volume from the hippocampal or caudate profile bilaterally. To ensure accurate volumetric analysis, the hippocampus was sampled by a magnification of 7px/mm3 in the axial view. The scale bar was used to correct for the magnification on the image in comparison to real measurements. The above procedure was also used for the caudate bilaterally, however, at a magnification of 8px/mm3 in the axial view. Thus, the stereological parameters for the volumetric estimation of both structures were optimised to achieve minimal errors in measurement. The schematic illustration of the Cavalieri principle as applied for MRI segmentation was shown (see Fig.1; Mayer et al. 2016).

***Volumetric Quantification using Fully-automatic (model-based) Segmentation with FSL-FIRST***

The fully automated segmentation of the volumes of the caudate and the hippocampus was carried out by employing FSL-FIRST (v.5.0.4, http://fsl.fmrib.ox.ac.uk/) after registration of images to a standard template (Center for Morphometric Analysis (CMA), MGH, Boston). These templates have been manually segmented as done previously (Filipek et al. 1994) to generate the models employed in FIRST. Following registration, the model-based approach using a Bayesian framework allows the probabilistic relationship of shape and intensity to be exploited in estimating volume (Patenaude 2007; Patenaude et al. 2011). To ensure best fit for each of the structures of interest, the empirically optimised default settings for each structure were applied. The accuracy of the segmentation was visually checked from the labels for the specific structures (L/R) in native space to ensure good registration. The diagrammatic illustration of the transformation of the images from native space into standard space for the segmentation and back to native space for volumetric analysis was shown (see Fig.3; Morey et al. 2009). Eight subjects were removed from all further analyses due to poor registration and volume extraction errors.

***Volumetric Quantification using Fully-automatic (patch-based) Segmentation with volBrain***

The volume of the caudate and the hippocampus was segmented by employing volBrain (v.1.0, <http://volbrain.upv.es>), another fully automated segmentation technique. The operational pipeline of this free online MRI brain volumetry system was described previously (Manjon and Coupe 2016). Briefly, this technique employs the multi-atlas patch-based label fusion segmentation technology (Coupé et al. 2011) to segment anatomical structures. The system firstly pre-processes the images through a spatially adaptive non-local means of denoising and corrects for intensity inhomogeneity roughly. The images are then registered to MNI space and again corrected for inhomogeneity to achieve a fine normalised intensity. The pre-processing sets the input image in the same geometric and intensity space. The volumetric estimation of the various brain structures at different scales are then achieved through a non-local intracranial cavity extraction and tissue classification. Furthermore, segmentation of the hemispheres and then the subcortical structures (including caudate and hippocampus) were achieved by distinguishing each pixel or patch through the Optimised PAtchMatch Label fusion (OPAL) strategy (Giraud et al, 2016; Bao and Chung 2016). Thus, subcortical labelling is based on a weighted label vote scheme through comparisons of pixels or patches. The image patch and the multi-atlas (manual labels) patches are compared after which weights are assigned according to patch similarity (Coupé et al. 2011; Giraud et al. 2016). This pipeline segments the hippocampus according to the EADC-ADNI Harmonised Protocol for Hippocampal Segmentation (Boccardi et al. 2015), the caudate was however segmented according to a local expert’s definition at the algorithm/pipeline development site. This processing pipeline was shown in a schematic diagram (see Fig.1; Manjon and Coupe 2016). A segmentation failure rate (0.4%, n=1) was reported and this was removed from all further analysis.

***Volumetric Quantification using Semi-automatic Segmentation with FreeSurfer***

The “recon-all” segmentation pipeline was employed for the delineation of both the caudate and the hippocampus bilaterally using FreeSurfer (v.5.1.0, http: //surfer. nmr.mgh.harvard.edu/). The technical operation of this free and documented volumetric image analysis suite was described previously (Fischl et al. 1999; Dale 1999, Fischl and Dale 2000; Fischl et al. 2002). Briefly, this technique segments by parcellating subcortical structures by affine registration to Talairach space depending on the differences in their voxel intensities. Thus, its estimations are based on the probability that each voxel will belong to a certain structure within the tissues of the brain. This pipeline is based on manual neuroanatomical segmentation protocols (Filipek et al. 1994) provided by the CMA. Schematic illustration of this outline was reported (see Fig. 3; Morey et al. 2009). The caudate and hippocampus volumes provided in the *aseg.stats* file were used in the analysis as they account for partial volume estimations and are deemed to be more accurate. Quality analyses included visual inspection of all the “recon-all” steps and the segmented structures as described in the established ENIGMA protocol (http://enigma.loni.ucla.edu/protocols/). Seventeen brain images were noted to have been inadequately skull stripped. These were corrected as described on the FreeSurfer website (https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki) by adjusting the watershed parameters using the [*recon-all -skullstrip -wsthresh XX -s XX*] and/or the *gcut* command without any further intervention manually. Three subjects were removed from all further analyses because of poor registration and extraction errors. The *aseg.mgz* volume files were converted from MGZ to NIFTI file format and transferred back into native space for visualisation.

***Statistical Analyses***

Statistical analyses were performed with volumes in native space using the Statistical Package for Social Sciences (SPSS) version 23.0 for Windows (SPSS Inc., IBM, New York, USA). The Shapiro–Wilk’s test was employed to test the normality of our data because of its relatively large power to detect normality (Razali and Wah 2011). In accordance with the comparative performance metrics outlined by Fenster and Chiu (2005), the degree to which the techniques agree with manual segmentation in accuracy, their level of consistency and efficiency were assessed.

***Analysis of Spatial Overlap and Volume difference***

These are commonly used in medical image segmentation, especially for comparing new techniques against known references (Taha and Hanbury 2015). As suggested for comparison of two techniques for inter-rater reliability by Collins et al. (1995), we investigated the level of agreement of each of the techniques with manual tracing in terms of their *percentage spatial overlap and volume difference* (Collins et al. 1995, Fischl et al. 2002; Han and Fischl 2007). The percent spatial overlap (Dice score) was estimated for the volumes relative to the gold standard volumes. For a structure **P**, segmented using two techniques, the percent spatial overlap is defined in equation (1). Thus, P1 is the volume as measured by the first technique and P2 is the volume from the second technique (gold standard). This was computed for each subject, structure and volumes compared to the gold standard.

**O(P1,P2) = [V(P1)** ∩ **V(P2)/V(P2)]\*100%** *equation (1)*

A maximum percent spatial overlap (100%) is achieved for identical volumes while less perfect spatial overlaps are assigned lower scores. Thus, slight shifts in the spatial location of a volume relative to another will cause a reduction in the overall percent spatial overlap between this volumes (Morey et al. 2009). Additionally, percent volume difference between volumes obtained with the other techniques and the gold standard were estimated as defined in equation (2). A positive volume difference indicates overestimation relative to the manually obtained volumes while a negative indicates underestimation and those approaching 0% indicate convergence of the volumes obtained from the other techniques and the gold standard.

**D(P1,P2) = [V(P1) – V(P2)/V(P2)]\*100%** *equation (2)*

We conducted a one-way ANOVA to compare the spatial overlaps and volumetric means obtained from the techniques. Tukey *post-hoc* test was conducted to indicate which of the techniques were significantly different from each other at 95% confidence interval (CI).

***Correlation Analysis between-techniques***

Although these volume comparisons are paired, some confounding factors are likely to affect the algorithms/pipelines of the techniques differently. For example, differences in age, gender and intracranial volume are reported to influence the effect a segmentation method has on statistical results (Barnes et al. 2010; Nordenskjold et al. 2015). Partial correlation analysis controlling for age and gender were therefore conducted to determine the level of consistency between manual tracing and the other segmentation techniques for the global sample. This statistical approach adjusts the crude correlation (Pearson’s) scores by eliminating the effect of the confounding variables. Intracranial volume was however not included as part of the control variables in this study as group difference effect was not the main subject under investigation and the inclusion of intracranial volume corrections would have excluded the computation of practical performance metrics such as the percentage volume overlap which depends on the actual position of voxels (Sanchez-Benavides et al. 2010). The magnitudes of these correlation coefficients were compared for existence of significant differences to assess performance with reference to the manual technique using the *Fisher r-to-z transformation* (Watson 2001). A total of six technique comparisons over the caudate and hippocampus were made to assess consistency bilaterally. An alpha adjustment at *p<0.01* for statistical significance in accordance to the *Bonferroni procedure* for a two-tailed hypothesis was employed. Furthermore, the partial correlation coefficients obtained from the techniques relative to manual tracing were compared between the patient and control group for existence of significant differences in consistency. These differences helped to establish consistency of the various techniques within different cohorts using the *Fisher r-to-z transformation* (Watson 2001).

Additionally, the intraclass correlation coefficient (ICC, single measures) was used to validate the absolute agreement between the techniques relative to the gold standard. A reliability score range of 0.75 - 0.90 is considered an optimum standard for adequate reproducibility and reliability (Hallgren 2012; Koo and Li 2016). The ICC was again used to determine the intra and inter-rater reliability between the raters for the manual and the stereological techniques.

***Analysis of potential technique biases: Bland – Altman plots***

The Bland-Altman plot is a popular statistical method for testing agreement of clinical techniques that measure continuous variables (Zaki et al. 2012). To further investigate the agreement and explore the extent of potential bias in the measurement of the volume of the subcortical structures by the techniques relative to the gold standard, we generated Bland-Altman plots (Bland and Altman, 1986, 1999). Contrary to the original assumptions for Bland-Altman plots, several recent studies (Schoemaker et al.2016; Makowski et al.2017), constructed the plots by computing the volumetric difference [V(P1) – V(P2)] relative to the manual and the other techniques for each structure per hemisphere against the corresponding manually traced volumes [V(P2)] as proposed by Krouwer (2008). A regression line was further integrated to the plot for each technique to assess potential biases in the estimation of the volumes and to observe whether the definition of the structures studied influenced the discrepancy between the gold standard and the other techniques. Additionally, all the integrated regression lines were tested for significance to confirm the absence of zero slopes.

**Results**

***Percentage spatial overlap between the manual and the other techniques***

As shown in Table 1 and Fig. 1a, There was a statistical significant difference between spatial overlap of the various segmentation techniques relative to manual segmentation for all regions (p<0.0001) [right caudate F (3, 1039) = 29.87, left caudate F (3, 1039) = 33.47; right hippocampus F (3, 1020) = 555.49 and left hippocampus F (3, 1020) = 555.49]. *Post-hoc* analysis further showed that the segmentation techniques were significantly different from each other (p<0.0001) in estimating the spatial locations of the left and right caudate except for comparisons between stereology and FSL-FIRST for the right (p=0.99) and left (p=0.13) caudate. The comparison between volBrain and FreeSurfer for the left caudate (p=0.09) were statistically not different. All comparisons demonstrated significant differences (p<0.0001) from each other in estimating the spatial location of the hippocampus bilaterally.

***Percentage volume difference between the manual and the other techniques***

The volume estimates by each technique for the caudate and the hippocampus are presented in Table 1 and Fig. 1b. Compared to manual estimates, all techniques studied underestimated (-16 to -24%) volume of the right and left caudate. In contrast, compared to manual estimates, the volume of the hippocampus was comparable when segmented using stereology (-7 to 5%) and was substantially overestimated by 44-75% using automated and semi-automated approaches. These mean volumetric differences were statistically significant at 95% confidence interval at p < 0.0001 for all regions [right caudate F(4,1289) = 168.92; left caudate F(4,1293) = 138.05; right hippocampus F(4,1263) = 1014.95 and left hippocampus F(4, 1258) = 843.10]. *Post-hoc* comparisons demonstrated for the left and right caudate that the segmentation techniques were significantly different from each other in estimating volume (p <0.0001) except for comparisons between stereology and FSL-FIRST (p = 0.99) and volBrain vs. FreeSurfer (p = 0.16) for the right caudate. Similarly, for the left caudate, stereology vs. FSL-FIRST (p = 0.45), stereology vs. volBrain (p = 0.08) and volBrain vs. FreeSurfer (p = 0.09) were not statistically different. All comparisons demonstrated significant differences from each other in estimating volumes of the hippocampus, with the least significant difference noted for comparisons between manual tracing and stereology for the left hippocampus (p = 0.02).

***Assessing consistency between the manual and the other techniques***

Each of the four techniques were highly correlated (0.77 – 0.87) with manual segmentation for the caudate bilaterally, and moderately correlated (0.35 - 0.62) for the hippocampus (Table 1, Fig. 2a, 2b). Absolute between-technique agreement (ICC) relative to manual volumes was generally fair (0.28 – 0.49) for the caudate. In contrast, hippocampal delineation by automated and semi-automated techniques gave poor (0.07 - 0.09) absolute agreement relative to manual segmentation, however, the stereologically acquired absolute hippocampal volumes agreed moderately (0.47 – 0.48).

The statistical comparison of the correlation coefficients of the global sample for each technique across the structures demonstrates that these are all not significantly different with the exception of FSL-FIRST vs. volBrain and FSL-FIRST vs. FreeSurfer for the caudate and hippocampus bilaterally (Table S1). In contrast, the comparison between stereology and FSL-FIRST for the left hippocampus was statistically significant (Table S1).

***Comparing correlations between the cohort groups***

The correlations obtained relative to the manual segmentation technique between patients and the healthy controls were compared between corresponding techniques to further assess consistency within the groups. The consistency of all the segmentation techniques within the groups was significantly not different from each other at estimating volumes except for stereology [left hippocampus (p = 0.008)], volBrain [right caudate (p = 0.003), left caudate (p = 0.005) and the left hippocampus (p =0.002) and FreeSurfer [left hippocampus (p = 0.025)] as shown in Table S2.

***Analysis of potential technique biases: Bland – Altman plots***

The volumetric bias estimations for the caudate and hippocampus were examined separately for each hemisphere. Compared to manual estimates, a relatively small bias range estimate was noted for the right (-1171.78 to -860.55) and left (-990.69 to -727.17) caudate. In contrast, a large bias range estimate was demonstrated for the right and left hippocampus; (-124.45 to 1902.38) and (187.56 to 1877.69) respectively. The regression analysis revealed a trend of significant negative bias in all the techniques across both structures within each hemisphere (Fig. 3a, 3b and Table 2).

***Consideration of Labour***

Manual segmentation was the most resource intensive of the techniques requiring 40 minutes per caudate and 60 minutes per hippocampus respectively for accurate bilateral segmentation. Stereological segmentation required approximately 50% of this time for each structure. FSL-FIRST required approximately 30% of the time of manual segmentation to segment the brain into 15 structures inclusive of both the caudate and hippocampus bilaterally. Furthermore, volBrain required approximately 15% of the time of manual segmentation to segment all the subcortical structures in addition to the ventricles regional macroscopic structures (hemispheres, cerebellum and brainstem) and volume of intracranial tissues (CSF, GM, WM). Finally, FreeSurfer required approximately 700% of the time of manual segmentation (12 hours) to segment the brain into 40 subcortical regions (including both the caudate and hippocampus) and other global brain measures such as intracranial volume. In addition, several output measures of FreeSurfer cortical reconstruction are generated during this time. FSL-FIRST and volBrain required no amount of manual intervention, however, FreeSurfer required 15 minutes per participant (including both the caudate and hippocampus) to undertake a thorough quality check of the outputs as per the standardised ENIGMA protocols, with an additional 30 minutes per participant required if they failed the skull strip test. Processing was entirely carried out on our system configuration: Intel(R) Xeon(R) CPU E5-2697 v3, 128GB @ 2.60GHz.

**Discussion**

The objective of this study was to compare the performance of stereological (Measure®), fully automated [model-based (FSL-FIRST v.5.0.4) and patch-based (volBrain v.1.0)] and semi–automated (FreeSurfer v.5.1) segmentation techniques to the current gold-standard method: manual segmentation (ITK-SNAP v.1.8) using representative brain structures, the caudate (*easy-to-segment)* and hippocampus (*difficult-to-segment*). The findings from this study demonstrate that automated techniques are fairly reliable in delineating the volumes of an *easy-to-segment* structure such as caudate in a relatively large cohort. Furthermore, we demonstrate that stereology shows high reliability in measuring both *easy-* and *difficult-to-segment* structures. Additionally, we found stereology to have a notably reduced labour intensity relative to manual segmentation whilst preserving the anatomical accuracy; a considerable merit of the manual approach. We further demonstrated, that volBrain was the preferred (Table 3) technique for segmentation of the caudate as it was least labour intensive and was fairly accurate anatomically. The extent to which each of these techniques rival the anatomical accuracy and efficiency of manual segmentation varies between caudate and hippocampus as detailed below. Taken together, our findings should help to inform other researchers about the considerable advantages and caveats (Table 3) of using each of these segmentation techniques in decision making for the now standard, large-scale brain morphometric studies ongoing in the field of neuropsychiatry.

Hippocampal volume estimates were anatomically accurate demonstrating an almost perfect spatial overlap (93 to 95%) and only slightly under/overestimated (-7 to 5%) using stereology relative to manual segmentation. Previous studies (Gundersen et al.1988; Yuen et al. 2001; García-Fiñana et al.2003) and the Measure® user guide indicated this volumetric difference to be acceptable especially in structures where greater variation is evident in the normal population. Furthermore, stereology demonstrated the highest observed absolute agreement (0.47-0.49) for estimation of the hippocampus. The hippocampus was significantly overestimated (44-75%) by the automated methods employed with spatial overlap of 57-69% recorded. The substantial overestimation of hippocampal volume observed is consistent with previous studies comparing both automated techniques to manual segmentation in adult (Doring et al. 2011; Cherbuin et al. 2009; Tae et al. 2008) and preadolescent (Schoemaker et al. 2016) populations. The trend of overestimation and poor reliability is however not peculiar to preadolescent populations but is similarly noted in adult populations, as previously reported. This study reported higher hippocampal volumes using the automated techniques [FreeSurfer (74-75%), volBrain (57-58%) and FSL-FIRST (44-47%)], partially as a result of a strict neuroanatomical definition of the structural boundaries for manual segmentation (Fig.3a). In particular, the alveus and fimbria were excluded, contrary to what was adopted by Schoemaker and colleagues demonstrating 28% (FSL-FIRST) and 60% (FreeSurfer) overestimations. Inclusion of these regions, together with anatomically mis-assigned surrounding grey and WM voxels (especially at the hippocampal head in some slices) and variance in the inclusion of the Andreas-Retzius and the Fasciolar gyrus from the tail region of the hippocampus across studies appear to account for the vast majority of the overestimation of volume that is evident when using the fully automated techniques. We noted (Fig. S3) the greatest amount of voxel misclassification using FreeSurfer, however application of manual interventions (skullstrip/control-point corrections) would likely significantly improve the quality of segmentation minimising these volume estimation errors (McCarthy et al. 2015).

Similarly, the even greater overestimation observed using semi-automated segmentation appears to be explained by the above factors with an added variable amount of parahippocampal and subicular/entorhinal cortices included in hippocampal volume estimation. The subiculum alone is reported to account for over 15% of the hippocampal volume when traced on an MRI (Geuze et al. 2005) thus its boundary definition will have a substantial effect on hippocampal volume estimation by all approaches and should be assessed carefully as a representative complex region in further development of automated methods. The lack of absolute agreement in volume estimation, although relevant anatomically may not always be essential in large scale neuropsychiatric research studies given the reasonable correlations with manual segmentation noted, which may suffice to detect case-control differences or disease progression markers.

In contrast, our representative *easy-to-segment* structure, the caudate was underestimated by all techniques (16-24%), however, a high (81-83%) spatial overlap was observed. The relatively uniform shape and strong contrast of the caudate from surrounding WM on MRI, allows its volume to be more susceptible to prior approximation by the rater using stereology. Volumetric estimation with stereology, however, is designed to be unbiased - without shape and volume assumptions of the structure under investigation (García-Fiñana et al.2003). The almost explicit prior information of the caudate (Fig. 4b) is suspected to have contributed to rater/methodological bias, hence the large discrepancy (21-24%) in the estimation of the volume of the caudate with stereology. The observed total variance (OCV2) in stereological estimation comprise of variance due to biological variability (CV2) and methodologically introduced variance (OCE2) (García-Fiñana et al.2003).

**OCV2 = CV2 + OCE2** *equation (3)*

Given the relatively low biological variance of the caudate, differences in volume estimation using stereology compared to manual segmentation may relate to methodological or rater bias. In contrast to the hippocampus, the relatively low MRI contrast and the more indistinct border (Fig.4a) has reduced its susceptibility to rater/methodological bias (Table 2). Our findings (-16 to -24%) of underestimation of the caudate is consistent and closely correspond in magnitude with previous studies utilising a semi-automated [-19.4% (Nazir et al.2014)] and fully-automated [-32.7 to 14.0% (Amann et al. 2015)] techniques; with volume underestimation here likely related to the exclusion of boundary voxels of the head of the caudate (labelled as nucleus accumbens or surrounding white matter) (Fig. 3b).

However, the correlation with manual segmentation was high for these techniques (r=0.77-0.87) as previously reported (Perlaki et al. 2017), although absolute agreement was relatively poor. Reasons for modest absolute agreement may be explained by neuroanatomical boundary definition differences derived from the different software’s employed, with varied sensitivity of tissue voxel classification algorithms (MRF, OPAL, Bayesian) potentially accounting for some of the observed variance.

Despite the underestimation of the caudate by FreeSurfer, it demonstrated a superior (0.45-0.49) absolute agreement relative to FIRST (0.28-0.30) and volBrain (0.37-0.45). This may potentially relate to the robustness of FreeSurfer’s segmentation pipeline adopted to assign unique labels to voxels for the spatial localisation of subcortical structures within a defined space based on MRF (Fischl et al. 2002). Furthermore, consistency of segmentation derived from volumes across a large sample will influence the sensitivity of the study to detect subtle differences between populations under study, and thus a consistent over/under-estimation with the segmentation techniques does not necessarily indicate a lack of validity (Schoemaker et al.2016).

To assess the reliability of each technique in terms of consistency of segmentation in the full sample, we calculated correlations between volumes derived from the techniques and manual segmentation volumes for both structures. Despite the under/overestimation of the volumes of the structures of interest relative to manual segmentation, the consistency of the techniques were relatively high for the caudate (0.77-0.87), and fair (FSL-FIRST: 0.35-0.42) or moderate (Stereology: 0.52-0.55, volBrain: 0.60-0.61, FreeSurfer: 0.56-0.62) for the hippocampus. Our findings extend previous works which assessed the operational consistency of automated techniques based on the correlation of volumes derived from automated and manual techniques in which a strong positive correlation indicated optimal consistency (Schoemaker et al.2016; Makowski et al.2017) otherwise, poor consistency or operational errors are assumed (Allen et al.2002) regardless of the absolute volume differences observed across approaches. Our findings of relatively low consistency of FSL-FIRST for hippocampal segmentation are consistent with previous studies (Morey et al. 2009; Schoemaker et al. 2016). The lower consistency of FSL-FIRST for hippocampal volume estimation may be attributable at least in part to differences in the sensitivity of operational algorithms at classifying voxels.

Within the diagnostic groups, the consistency of all techniques at segmenting the caudate for the patient and the control group was high, however, a decreasing trend of consistency was observed, particularly for the healthy control group when segmenting the hippocampus (Table 1) as reported previously (Sanchez-Benavides et al. 2010). Furthermore, the correlation coefficients within the diagnostic groups were compared for significant differences to assess consistency of each of the techniques within the groups. Significant differences were noted between techniques when segmenting the left hippocampus with further differences in consistency noted with volBrain for all structures. This potentially suggests volBrain was particularly sensitive at detecting neuroanatomical differences between groups, however the patient group was heterogeneous in nature and caution is thus required with such an interpretation. Thus, we suggest further that the differing rates/levels and effect size of atrophy (neurodegenerative) and malformations (neurodevelopmental) in brain pathologies together with other essential factors (Table 3) must be thoroughly considered when choosing a segmentation technique especially for processing of brain MRI of clinical groups.

The Bland–Altman plots demonstrated relatively few difference points (Fig. 3a, 3b) outside of the established limits (± 1.96SD), with regards to the caudate and hippocampal volume bilaterally. Most of these difference points represented images for similar subjects across all techniques. This suggests that operational or systematic errors did not unduly influence segmentation and these subjects were potential outliers. The plots further suggested that biases existed for all the techniques (Fig. 3a, 3b and Table 2) relative to manual segmentation across all the structures. The greatest and least biases were observed with stereology and FreeSurfer when segmenting the caudate respectively. The direct reverse of this trend was observed when segmenting the hippocampus. Furthermore, the integrated regression plots demonstrated a significant negative correlation across all the structures with all techniques relative to manual segmentation suggesting a trend of relatively smaller structures being more susceptible to volume overestimation, a finding consistent with some previous studies (Makowski et al. 2017; Schoemaker et al. 2016).

Manual tracing is the most labour intensive approach and consequently in large-scale studies where an *easy-to-segment* structure is being examined, automated techniques demonstrate clear advantages. Careful consideration of a wide range of factors including anatomical accuracy, availability of expertise and time, consistency and the number of structures being estimated should, we suggest influence the choice of segmentation technique employed (Table 3).

There are a number of limitations associated with this study. Firstly, only two structures were examined; however given that these two structures are implicated in several neuropsychiatric disorders and are at either end of the spectrum in relation to the ease of segmentaiton, findings in relation to these structures are likely generalisable to several other brain structures. Secondly, data was acquired using a 1.5T MR scanner, and it is possible that automated segmentation may perform more accurately on images acquired with a 3T MR scanner. Thirdly, absolute volumes were used for this study rather than intracranial-corrected volumes however, group difference effects were not under examination in this study and the inclusion of intracranial volume corrections would have excluded the computation of practical performance metrics such as the percentage volume overlap which depends on the actual position of voxels (Sanchez-Benavides et al. 2010). This study has several strengths including the comparison of a range of different segmentation techniques (stereology, semi-automated and two automated) with quantified bias estimates across structures, a relatively large sample size and the inclusion of both healthy controls and individuals with significant neuropsychiatric disorders. Thus these findings are likely generalisable to large-scale global collaborations using semi-automated techniques such as ENIGMA, the Human Connectome Project and the Alzheimer's Disease Neuroimaging Initiative (ADNI) for example (Hibar et al.2016; van Erp et al.2016; Franke et al.2016; Schuff et al.2009) which have investigated to date the influence of various demographic and clinical factors on sub/cortical structures.

**Conclusion**

Our findings indicate that the anatomical accuracy and consistency of subcortical segmentation is dependent upon the anatomical location and adjacency to other grey matter structures which are difficult to differentiate from the structure in question. Furthermore, volumetric estimates from automated techniques are not always comparable to those obtained from manual segmentation especially when examining complex anatomical structures. Consequently, caution is required in interpretation of volume estimates depending on the structure being examined and the segmentation technique employed. Thus, recognising an optimum segmentation algorithm/technique during the design phase of a study is important as a single approach is not necessarily perfect at segmenting all structures. In a typical dataset of MRI images used in neuropsychiatric studies, standard automated segmentation techniques appear to provide reasonable accuracy for an easy-to-segment structure such as the caudate; however, for a more difficult to segment structure such as the hippocampus, automated techniques provide reasonable correlation with volume but poor absolute approximation. This indicates that manual or stereological volume estimation should be considered for studies that require high levels of precision such as those with small sample sizes, particularly if difficult to segment structures are being examined. A wide range of factors including the number of structures being examined and the availability of expertise and time should additionally be considered prior to the choice of segmentation technique in neuropsychiatric research studies.

**Conflict of Interest** The authors declare that they have no conflict of interest.

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent** All participants provided written informed consent for the relevant studies.

**References**

Ahmed, M., Cannon, D. M., Scanlon, C., Holleran, L., Schmidt, H., McFarland, J., et al. (2015). Progressive Brain Atrophy and Cortical Thinning in Schizophrenia after Commencing Clozapine Treatment. Neuropsychopharmacology, 40(10), 2409-2417, doi:10.1038/npp.2015.90.

Allen, J. S., Damasio, H., & Grabowski, T. J. (2002). Normal neuroanatomical variation in the human brain: an MRI-volumetric study. Am J Phys Anthropol, 118(4), 341-358, doi:10.1002/ajpa.10092.

Altshuler, L. L., Bartzokis, G., Grieder, T., Curran, J., Jimenez, T., Leight, K., et al. (2000). An MRI study of temporal lobe structures in men with bipolar disorder or schizophrenia. Biol Psychiatry, 48(2), 147-162.

Altshuler, L. L., Bartzokis, G., Grieder, T., Curran, J., & Mintz, J. (1998). Amygdala enlargement in bipolar disorder and hippocampal reduction in schizophrenia: an MRI study demonstrating neuroanatomic specificity. Arch Gen Psychiatry, 55(7), 663-664.

Amann, M., Andělová, M., Pfister, A., Mueller-Lenke, N., Traud, S., Reinhardt, J., et al. (2015). Subcortical brain segmentation of two dimensional T1-weighted data sets with FMRIB's Integrated Registration and Segmentation Tool (FIRST). NeuroImage: Clinical, 7, 43-52, doi:http://dx.doi.org/10.1016/j.nicl.2014.11.010.

Bao, S., & Chung, A. C. S. (2017). Feature Sensitive Label Fusion with Random Walker for Atlas-Based Image Segmentation. IEEE Transactions on Image Processing, 26(6), 2797-2810, doi:10.1109/TIP.2017.2691799.

Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M. D., Lehmann, M., Hobbs, N., et al. (2010). Head size, age and gender adjustment in MRI studies: a necessary nuisance? NeuroImage, 53(4), 1244-1255, doi:https://doi.org/10.1016/j.neuroimage.2010.06.025.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. Statistical Methods in Medical Research, 8(2), 135-160, doi: 10.1177/096228029900800204.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307-310, doi:http://dx.doi.org/10.1016/S0140-6736 (86)90837-8.

Boccardi, M., Bocchetta, M., Apostolova, L. G., Barnes, J., Bartzokis, G., Corbetta, G., et al. (2015). Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimer's & Dementia, 11(2), 126-138, doi:https://doi.org/10.1016/j.jalz.2014.02.009.

Cahn, W., Pol, H., Lems, E. E., & et al. (2002). Brain volume changes in first-episode schizophrenia: A 1-year follow-up study. Arch Gen Psychiatry, 59(11), 1002-1010, doi:10.1001/archpsyc.59.11.1002.

Cherbuin, N., Anstey, K. J., Réglade-Meslin, C., & Sachdev, P. S. (2009). In Vivo Hippocampal Measurement and Memory: A Comparison of Manual Tracing and Automated Segmentation in a Large Community-Based Sample. PLoS ONE, 4(4), e5265, doi:10.1371/journal.pone.0005265.

Collins, D. L., Holmes, C. J., Peters, T. M., & Evans, A. C. (1995). Automatic 3-D model-based neuroanatomical segmentation. Hum Brain Mapp, 3(3), 190-208, doi:10.1002/hbm.460030304.

Coupé, P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., & Collins, D. L. (2011). Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage, 54(2), 940-954, doi:http://dx.doi.org/10.1016/j.neuroimage.2010.09.018.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage, 9(2), 179-194, doi:10.1006/nimg.1998.0395.

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

Emsell, L., Langan, C., Van Hecke, W., Barker, G. J., Leemans, A., Sunaert, S., et al. (2013). White matter differences in euthymic bipolar I disorder: a combined magnetic resonance imaging and diffusion tensor imaging voxel-based study. Bipolar Disord, 15(4), 365-376, doi:10.1111/bdi.12073.

Ertekin, T., Acer, N., İçer, S., Vurdem, Ü. E., Çınar, Ş., & Özçelik, Ö. (2015). Volume estimation of the subcortical structures in Parkinson's disease using magnetic resonance imaging: A methodological study. [Article]. Neurology Asia, 20(2), 143-153.

Fenster, A., & Chiu, B. (2005). Evaluation of Segmentation algorithms for Medical Imaging. Conf Proc IEEE Eng Med Biol Soc, 7, 7186-7189, doi:10.1109/iembs.2005.1616166.

Filipek, P. A., Richelme, C., Kennedy, D. N., & Caviness, V. S., Jr. (1994). The young adult human brain: an MRI-based morphometric analysis. Cereb Cortex, 4(4), 344-360.

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci U S A, 97(20), 11050-11055, doi:10.1073/pnas.200033797.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron, 33(3), 341-355.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. NeuroImage, 9(2), 195-207, doi:10.1006/nimg.1998.0396.

Garcia, Y., Breen, A., Burugapalli, K., Dockery, P., & Pandit, A. (2007). Stereological methods to assess tissue response for tissue-engineered scaffolds. Biomaterials, 28(2), 175-186, doi:http://dx.doi.org/10.1016/j.biomaterials.2006.08.037.

García-Fiñana, M., Cruz-Orive, L. M., Mackay, C. E., Pakkenberg, B., & Roberts, N. (2003). Comparison of MR imaging against physical sectioning to estimate the volume of human cerebral compartments. NeuroImage, 18(2), 505-516, doi:http://dx.doi.org/10.1016/S1053-8119 (02)00021-6.

Geuze, E., Vermetten, E., & Bremner, J. D. (2005). MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed. Mol Psychiatry, 10(2), 147-159, doi:10.1038/sj.mp.4001580.

Giraud, R., Ta, V.-T., Papadakis, N., Manjón, J. V., Collins, D. L., & Coupé, P. (2016). An Optimized PatchMatch for multi-scale and multi-feature label fusion. NeuroImage, 124, 770-782, doi:http://dx.doi.org/10.1016/j.neuroimage.2015.07.076.

Grimm, O., Pohlack, S., Cacciaglia, R., Winkelmann, T., Plichta, M. M., Demirakca, T., et al. (2015). Amygdalar and hippocampal volume: A comparison between manual segmentation, Freesurfer and VBM. Journal of Neuroscience Methods, 253, 254-261, doi:http://dx.doi.org/10.1016/j.jneumeth.2015.05.024.

Gundersen, H. J., Bagger, P., Bendtsen, T. F., Evans, S. M., Korbo, L., Marcussen, N., et al. (1988). The new stereological tools: disector, fractionator, nucleator and point sampled intercepts and their use in pathological research and diagnosis. APMIS, 96(10), 857-881.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in quantitative methods for psychology, 8(1), 23-34.

Han, X., & Fischl, B. (2007). Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE Trans Med Imaging, 26(4), 479-486, doi:10.1109/tmi.2007.893282.

Hibar, D. P., Westlye, L. T., van Erp, T. G. M., Rasmussen, J., Leonardo, C. D., Faskowitz, J., et al. (2016). Subcortical volumetric abnormalities in bipolar disorder. [Original Article]. Mol Psychiatry, 21(12), 1710-1716, doi:10.1038/mp.2015.227.

Kenney, J., Anderson-Schmidt, H., Scanlon, C., Arndt, S., Scherz, E., McInerney, S., et al. (2015). Cognitive course in first-episode psychosis and clinical correlates: A 4 year longitudinal study using the MATRICS Consensus Cognitive Battery. Schizophr Res, 169(1-3), 101-108, doi:10.1016/j.schres.2015.09.007.

Keshavan, M. S., Anderson, S., Beckwith, C., Nash, K., Pettegrew, J. W., & Krishnan, K. R. R. (1995). A comparison of stereology and segmentation techniques for volumetric measurements of lateral ventricles in magnetic resonance imaging. Psychiatry Research: Neuroimaging, 61(1), 53-60, doi:http://dx.doi.org/10.1016/0925-4927 (95)02446-5.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine, 15(2), 155-163, doi:https://doi.org/10.1016/j.jcm.2016.02.012.

Krouwer, J. S. (2008). Why Bland–Altman plots should use X, not (Y+X)/2 when X is a reference method. Statistics in Medicine, 27(5), 778-780, doi:10.1002/sim.3086.

Looi, J. C., Lindberg, O., Liberg, B., Tatham, V., Kumar, R., Maller, J., et al. (2008). Volumetrics of the caudate nucleus: reliability and validity of a new manual tracing protocol. Psychiatry Res, 163(3), 279-288, doi:10.1016/j.pscychresns.2007.07.005.

Makowski, C., Béland, S., Kostopoulos, P., Bhagwat, N., Devenyi, G. A., Malla, A. K., et al. (2017). Evaluating accuracy of striatal, pallidal, and thalamic segmentation methods: Comparing automated approaches to manual delineation. NeuroImage, doi:http://dx.doi.org/10.1016/j.neuroimage.2017.02.069.

Mamah, D., Alpert, K. I., Barch, D. M., Csernansky, J. G., & Wang, L. (2016). Subcortical neuromorphometry in schizophrenia spectrum and bipolar disorders. NeuroImage: Clinical, 11, 276-286, doi:10.1016/j.nicl.2016.02.011.

Mamah, D., Harms, M. P., Barch, D., Styner, M., Lieberman, J. A., & Wang, L. (2012). Hippocampal Shape and Volume Changes with Antipsychotics in Early Stage Psychotic Illness. Frontiers in Psychiatry, 3, 96, doi:10.3389/fpsyt.2012.00096.

Manjón, J. V., & Coupé, P. (2016). volBrain: An Online MRI Brain Volumetry System. Frontiers in Neuroinformatics, 10, 30, doi:10.3389/fninf.2016.00030.

Mayer, K. N., Latal, B., Knirsch, W., Scheer, I., von Rhein, M., Reich, B., et al. (2016). Comparison of automated brain volumetry methods with stereology in children aged 2 to 3 years. [journal article]. Neuroradiology, 58(9), 901-910, doi:10.1007/s00234-016-1714-x.

McCarthy, C. S., Ramprashad, A., Thompson, C., Botti, J.-A., Coman, I. L., & Kates, W. R. (2015). A comparison of FreeSurfer-generated data with and without manual intervention. [Original Research]. Frontiers in Neuroscience, 9(379), doi:10.3389/fnins.2015.00379.

McFarland, J., Cannon, D. M., Schmidt, H., Ahmed, M., Hehir, S., Emsell, L., et al. (2013). Association of grey matter volume deviation with insight impairment in first-episode affective and non-affective psychosis. [journal article]. European Archives of Psychiatry and Clinical Neuroscience, 263(2), 133-141, doi: 10.1007/s00406-012-0333-8.

Morey, R. A., Petty, C. M., Xu, Y., Pannu Hayes, J., Wagner, H. R., Lewis, D. V., et al. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. NeuroImage, 45(3), 855-866, doi:http://dx.doi.org/10.1016/j.neuroimage.2008.12.033.

Nazir, M., Cleret de Langavant, L., Brugieres, P., Gaura, V., Lavisse, S., Youssov, K., Bachoud-Levi, A.-C., Remy, P. Comparison of three techniques to measure longitudinally striatal volume in Huntington's disease patients [[abstract]]. Movement Disorders 29(Supple 1), 227.

Nordenskjöld, R., Malmberg, F., Larsson, E.-M., Simmons, A., Ahlström, H., Johansson, L., et al. (2015). Intracranial volume normalization methods: Considerations when investigating gender differences in regional brain volume. Psychiatry Research: Neuroimaging, 231(3), 227-235, doi:https://doi.org/10.1016/j.pscychresns.2014.11.011.

Okada, N., Fukunaga, M., Yamashita, F., Koshiyama, D., Yamamori, H., Ohi, K., et al. (2016). Abnormal asymmetries in subcortical brain volume in schizophrenia. [Original Article]. Mol Psychiatry, 21(10), 1460-1466, doi:10.1038/mp.2015.209.

Pardoe, H. R., Pell, G. S., Abbott, D. F., & Jackson, G. D. (2009). Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? Epilepsia, 50(12), 2586-2592.

Patenaude, B., Smith, S., Kennedy, D., & Jenkinson, M. (2007). Bayesian shape and appearance models, Technical report TR07BP1, FMRIB Centre - University of Oxford

Patenaude, B., Smith, S. M., Kennedy, D. N., & Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage, 56(3), 907-922, doi:10.1016/j.neuroimage.2011.02.046.

Perlaki, G., Horvath, R., Nagy, S. A., Bogner, P., Doczi, T., Janszky, J., et al. (2017). Comparison of accuracy between FSL’s FIRST and Freesurfer for caudate nucleus and putamen segmentation. Scientific Reports, 7, 2418, doi: 10.1038/s41598-017-02584-5.

Quigley, S. J., Scanlon, C., Kilmartin, L., Emsell, L., Langan, C., Hallahan, B., et al. (2015). Volume and shape analysis of subcortical brain structures and ventricles in euthymic bipolar I disorder. Psychiatry Research: Neuroimaging, 233(3), 324-330, doi:http://dx.doi.org/10.1016/j.pscychresns.2015.05.012.

Razali, N. M. & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33.

Renteria, M. E., Schmaal, L., Hibar, D. P., Couvy-Duchesne, B., Strike, L. T., Mills, N. T., et al. (2017). Subcortical brain structure and suicidal behaviour in major depressive disorder: a meta-analysis from the ENIGMA-MDD working group. [Original Article]. Transl Psychiatry, 7, e1116, doi:10.1038/tp.2017.84.

Rodionov, R., Chupin, M., Williams, E., Hammers, A., Kesavadas, C., & Lemieux, L. (2009). Evaluation of atlas-based segmentation of hippocampi in healthy humans. Magn Reson Imaging, 27(8), 1104-1109, doi:10.1016/j.mri.2009.01.008.

Sacchet, M. D., Livermore, E. E., Iglesias, J. E., Glover, G. H., & Gotlib, I. H. (2015). Subcortical volumes differentiate Major Depressive Disorder, Bipolar Disorder, and remitted Major Depressive Disorder. Journal of Psychiatric Research, 68, 91-98, doi:http://dx.doi.org/10.1016/j.jpsychires.2015.06.002.

Sánchez-Benavides, G., Gómez-Ansón, B., Sainz, A., Vives, Y., Delfino, M., & Peña-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(3), 219-225, doi:http://dx.doi.org/10.1016/j.pscychresns.2009.10.011.

Scanlon, C., Anderson-Schmidt, H., Kilmartin, L., McInerney, S., Kenney, J., McFarland, J., et al. (2014). Cortical thinning and caudate abnormalities in first episode psychosis and their association with clinical outcome. Schizophrenia Research, 159(1), 36-42, doi:http://dx.doi.org/10.1016/j.schres.2014.07.030.

Schoemaker, D., Buss, C., Head, K., Sandman, C. A., Davis, E. P., Chakravarty, M. M., et al. (2016). Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. NeuroImage, 129, 1-14, doi:http://dx.doi.org/10.1016/j.neuroimage.2016.01.038.

Sheline, Y. I., Sanghavi, M., Mintun, M. A., & Gado, M. H. (1999). Depression Duration But Not Age Predicts Hippocampal Volume Loss in Medically Healthy Women with Recurrent Major Depression. The Journal of Neuroscience, 19(12), 5034-5043.

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging, 17(1), 87-97, doi:10.1109/42.668698.

Strakowski, S. M., DelBello, M. P., Sax, K. W., & et al. (1999). Brain magnetic resonance imaging of structural abnormalities in bipolar disorder. Arch Gen Psychiatry, 56(3), 254-260, doi:10.1001/archpsyc.56.3.254.

Tae, W. S., Kim, S. S., Lee, K. U., Nam, E.-C., & Kim, K. W. (2008). Validation of hippocampal volumes measured using a manual method and two automated methods (FreeSurfer and IBASPM) in chronic major depressive disorder. [journal article]. Neuroradiology, 50(7), 569, doi: 10.1007/s00234-008-0383-9.

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. [journal article]. BMC Medical Imaging, 15(1), 29, doi: 10.1186/s12880-015-0068-x.

van Erp, T. G., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., et al. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. Mol Psychiatry, 21(4), 547-553, doi:10.1038/mp.2015.63.

Velakoulis, D., Wood, S. J., Wong, M. T., McGorry, P. D., Yung, A., Phillips, L., et al. (2006). Hippocampal and amygdala volumes according to psychosis stage and diagnosis: a magnetic resonance imaging study of chronic schizophrenia, first-episode psychosis, and ultra-high-risk individuals. Arch Gen Psychiatry, 63(2), 139-149, doi:10.1001/archpsyc.63.2.139.

Watson, R. (2001). SPSS Survival Manual by Julie Pallant, Open University Press, Buckingham, 2001, 286 pages, ISBN 0 335 20890 8. Journal of Advanced Nursing, 36(3), 478-478, doi:10.1046/j.1365-2648.2001.2027c.x.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage, 31(3), 1116-1128, doi:10.1016/j.neuroimage.2006.01.015.

Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N. A. (2012). Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. PLoS ONE, 7(5), e37908, doi: 10.1371/journal.pone.0037908

**Figures**

**Fig.** 1a *Percentage volume overlap and standard deviations showing differences in estimating the spatial locations between techniques across structures*



**Fig.** 1a*Significant differences (\*) were observed among the techniques relative to manual tracing across each structure. Tukey’s post-hoc analysis revealed significant differences for all comparisons except for those**marked by asterisks (\*\*)*

**Fig.** 1b *Mean volumes and standard deviations showing volume differences between techniques across structures*



**Fig.** 1b*Significant differences (\*) were observed among the techniques across each structure. Tukey’s post-hoc analysis revealed significant differences for all comparisons except for those**marked by asterisks (\*\*) and least significance noted (†)*

|  |  |  |
| --- | --- | --- |
|  | **Right Caudate (mean volume mm3)** | **Right Hippocampus (mean volume mm3)** |
| **Stereology**  **(mean volume mm3)** |  |  |
|  | **Right Caudate (mean volume mm3)** | **Right Hippocampus (mean volume mm3)** |
| **FSL-FIRST**  **(mean volume mm3)** |  |  |

**Fig.** 2a *Scatterplots demonstrating values as assessed by manual segmentation in comparison to other techniques to assess consistency in the segmentation of the right caudate and hippocampus.*

**Fig.** 2a*The statistical values corresponding to the above correlations are presented in Table 1. The corresponding graphs for the left caudate and hippocampus are provided in the supplementary material (Fig.S1)*

|  |  |  |
| --- | --- | --- |
|  | **Right Caudate (mean volume mm3)** | **Right Hippocampus (mean volume mm3)** |
| **volBrain**  **(mean volume mm3)** |  |  |
|  | **Right Caudate (mean volume mm3)** | **Right Hippocampus (mean volume mm3)** |
| **FreeSurfer**  **(mean volume mm3)** |  |  |

**Fig.** 2b*Scatterplots demonstrating values as assessed by manual segmentation in comparison to other techniques to assess consistency in the segmentation of the right caudate and hippocampus*

**Fig.** 2b *The statistical values corresponding to the above correlations are presented in Table 1. The corresponding graphs for the left caudate and hippocampus are provided in the supplementary material (Fig.S2)*

**Fig.** 3a *Comparison of the segmented output (right hippocampus) from the techniques*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Native | Manual | FSL | volBrain | FreeSurfer |
| Axial  (z=65) | C:\Users\Administrator\Desktop\SegStudyImages\Manual\hippocampus\native_hippo_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\Manual\hippocampus\manual_hippo_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\FSL\Hippocampus\FSL_hippo_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\volBrain\hippocampus\volBrain_hippo_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\FreeSurfer\FreeSurfer_hippo_axial.png |
| Coronal  (y =240) | C:\Users\Administrator\Desktop\SegStudyImages\Manual\hippocampus\native_hippo_cor.png | C:\Users\Administrator\Desktop\SegStudyImages\Manual\hippocampus\manual_hippo_cor.png | C:\Users\Administrator\Desktop\SegStudyImages\FSL\Hippocampus\FSL_hippo_cor.png | C:\Users\Administrator\Desktop\SegStudyImages\volBrain\hippocampus\volBrain_hippo_cor.png | C:\Users\Administrator\Desktop\SegStudyImages\FreeSurfer\FreeSurfer_hippo_cor.png |
| Sagittal  (x=317) | C:\Users\Administrator\Desktop\SegStudyImages\Manual\hippocampus\native_hippo_sag.png | C:\Users\Administrator\Desktop\SegStudyImages\Manual\hippocampus\manual_hippo_sag.png | C:\Users\Administrator\Desktop\SegStudyImages\FSL\Hippocampus\FSL_hippo_sag.png | C:\Users\Administrator\Desktop\SegStudyImages\volBrain\hippocampus\volBrain_hippo_sag.png | C:\Users\Administrator\Desktop\SegStudyImages\FreeSurfer\FreeSurfer_hippo_sag.png |

**Fig.** 3a *Images utilising Measure (Stereology) were unfortunately unavailable.*

**Fig.** 3b *Comparison of the segmented output (right caudate) from the techniques*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Native | Manual | FSL | volBrain | FreeSurfer |
| Axial  (z=83) | C:\Users\Administrator\Desktop\SegStudyImages\Manual\Caudate\caudate_native_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\Manual\Caudate\manual_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\FSL\Caudate\FSL_Caudate_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\volBrain\volBrain_Caudate_axial.png | C:\Users\Administrator\Desktop\SegStudyImages\FreeSurfer\Caudate\FreeSurfer_Caudate_axial.png |
| Coronal  (y =288) | C:\Users\Administrator\Desktop\SegStudyImages\Manual\Caudate\caudate_native_cor.png | C:\Users\Administrator\Desktop\SegStudyImages\Manual\Caudate\manual_cor.png | C:\Users\Administrator\Desktop\SegStudyImages\FSL\Caudate\FSL_Caudate_coronal.png | C:\Users\Administrator\Desktop\SegStudyImages\volBrain\volBrain_Caudate_cor.png | C:\Users\Administrator\Desktop\SegStudyImages\FreeSurfer\Caudate\FreeSurfer_Caudate_cor.png |
| Sagittal  (x=277) | C:\Users\Administrator\Desktop\SegStudyImages\Manual\Caudate\caudate_native_sag.png | C:\Users\Administrator\Desktop\SegStudyImages\Manual\Caudate\manual_sag.png | C:\Users\Administrator\Desktop\SegStudyImages\FSL\Caudate\FSL_Caudate_sagittal.png | C:\Users\Administrator\Desktop\SegStudyImages\volBrain\volBrain_Caudate_sag.png | C:\Users\Administrator\Desktop\SegStudyImages\FreeSurfer\Caudate\FreeSurfer_Caudate_sag.png |

**Fig.** 3b *Images utilising Measure (Stereology) were unfortunately unavailable.*

**Fig.** 4a Right Caudate *Bland-Altman plots for bias estimation between manual segmentation and* ***A:*** *Stereology,* ***B:*** *FSL,* ***C:*** *volBrain and* ***D:*** *FreeSurfer*

|  |  |
| --- | --- |
|  |  |
|  |  |

**Fig.** 4a *The red regression lines were fit to show potential bias in volume estimation, the mean is represented with the continuous line while the lower (-1.96 x standard deviation) and upper limits (+1.96 x standard deviation) of agreement are represented with broken lines. The graph for the left caudate is provided in the supplementary material (Fig.S4)*

**Fig.** 4b Right Hippocampus **–** *Bland-Altman plots for bias estimation between manual segmentation and* ***A:*** *Stereology,* ***B:*** *FSL,* ***C:*** *volBrain and* ***D:*** *FreeSurfer*

|  |  |
| --- | --- |
|  |  |
|  |  |

**Fig.** 4b*The red regression lines were fit to show potential bias in volume estimation, the mean is represented with the continuous line while the lower (-1.96 x standard deviation) and upper limits (+1.96 x standard deviation) of agreement are represented with broken lines.**The graph for the left hippocampus is provided in the supplementary material (Fig.S5)*

**Tables**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Anatomy/Technique** | **Structure Volume**  (Mean mm3  ± SD) | **Comparison of techniques to manual tracing** | | | | **Diagnostic Groups**  (*Partial Correlation)* | |
| *%Volume*  *Difference*  *± SD* | *%Volume Overlap± SD* | *Intraclass Correlation Coefficient*  *( 95% CI[Lower –Upper])* | *Partial Correlation* ***(n=281)***  *(r, p)* |
| *Healthy Controls*  ***(n=104)***  *(r, p)* | *Patients*  ***(n=177)***  *(r, p)* |
| ***Right Caudate*** | | | | | | | |
| ITK-SNAP | **4836±666** |  |  |  |  |  |  |
| Stereology | 3691±584 | -24±6 | 81±4 | 0.30 [-0.03-0.67] | 0.83, <0.0001 | 0.81, <0.0001 | 0.85, <0.0001 |
| FSL-FIRST | 3716±490 | -23±7 | 81±5 | 0.28 [-0.05-0.64] | 0.77, <0.0001 | 0.79, <0.0001 | 0.75, <0.0001 |
| volBrain | 3869±525 | -20±6 | 83±6 | 0.37 [-0.05-0.73] | 0.86, <0.0001 | 0.92, <0.0001 | 0.83, <0.0001 |
| FreeSurfer | 3978±590 | -18±7 | 85±7 | 0.45 [-0.06-0.79] | 0.85, <0.0001 | 0.89, <0.0001 | 0.83, <0.0001 |
| ***Left Caudate*** | | | | | | | |
| ITK-SNAP | **4660±643** |  |  |  |  |  |  |
| Stereology | 3694±579 | -21±6 | 83±4 | 0.36 [-0.04-0.73] | 0.83, <0.0001 | 0.78, <0.0001 | 0.86, <0.0001 |
| FSL-FIRST | 3608±469 | -23±7 | 82±5 | 0.30 [-0.05-0.65] | 0.77, <0.0001 | 0.77, <0.0001 | 0.78, <0.0001 |
| volBrain | 3822±546 | -20±6 | 85±5 | 0.45 [-0.05-0.79] | 0.87, <0.0001 | 0.92, <0.0001 | 0.87, <0.0001 |
| FreeSurfer | 3920±582 | -16±6 | 86±6 | 0.49 [-0.07-0.81] | 0.85, <0.0001 | 0.87, <0.0001 | 0.85, <0.0001 |
| ***Right Hippocampus*** | | | | | | | |
| ITK-SNAP | **2545±533** |  |  |  |  |  |  |
| Stereology | 2366±366 | -7±18 | 93±7 | 0.49 [0.37-0.59] | 0.52, <0.0001 | 0.38, <0.0001 | 0.56, <0.0001 |
| FSL-FIRST | 3732±445 | 47±30 | 68±13 | 0.10 [-0.06-0.30] | 0.42, <0.0001 | 0.45, <0.0001 | 0.40, <0.0001 |
| volBrain | 4019±471 | 58±29 | 63±11 | 0.10 [-0.04-0.33] | 0.61, <0.0001 | 0.52, <0.0001 | 0.67, <0.0001 |
| FreeSurfer | 4454±445 | 75±30 | 57±9 | 0.08 [-0.02-0.29] | 0.62, <0.0001 | 0.56, <0.0001 | 0.63, <0.0001 |
| ***Left Hippocampus*** | | | | | | | |
| ITK-SNAP | **2526±515** |  |  |  |  |  |  |
| Stereology | 2654±342 | 5±19 | 95±10 | 0.47 [0.30-0.60] | 0.55, <0.0001 | 0.31, <0.001 | 0.61, <0.0001 |
| FSL-FIRST | 3649±432 | 44±31 | 69±13 | 0.08 [-0.06-0.26] | 0.35, <0.0001 | 0.25, <0.0001 | 0.38, <0.0001 |
| volBrain | 3954±469 | 57±29 | 64±11 | 0.09 [-0.04-0.32] | 0.60, <0.0001 | 0.40, <0.0001 | 0.69, <0.0001 |
| FreeSurfer | 4406±451 | 74±30 | 57±9 | 0.07 [-0.02-0.25] | 0.56, <0.0001 | 0.38, <0.0001 | 0.61, <0.0001 |

**Table 1** *Comparison of relative and absolute volumes as derived from stereology, fully-automated and semi-automated techniques against manual tracing*

**Table** 1 *All volumes compared are in native space. For the partial correlation, age and gender were used as the co-variate*

**Table 2** *Quantitative bias estimation and regression analysis**of techniques across structures*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Anatomy/Structure*** | ***Estimation of Bias*** | | ***Regression Analysis*** | |
| ***Bias (mm3)*** | ***SD*** | ***r*** | ***p*** |
| **Right Caudate** | | | | |
| *Stereology* | -1171.78 | 317.64 | -0.39 | <0.0001 |
| *FSL-FIRST* | -1049.23 | 368.57 | -0.62 | <0.0001 |
| *volBrain* | -961.81 | 313.36 | -0.62 | <0.0001 |
| *FreeSurfer* | -860.55 | 325.29 | -0.45 | <0.0001 |
| **Left Caudate** | | | | |
| *Stereology* | -990.69 | 304.63 | -0.35 | <0.0001 |
| *FSL-FIRST* | -980.41 | 349.56 | -0.62 | <0.0001 |
| *volBrain* | -821.20 | 282.98 | -0.57 | <0.0001 |
| *FreeSurfer* | -727.17 | 297.46 | -0.48 | <0.0001 |
| **Right Hippocampus** | | | | |
| *Stereology* | -124.45 | 447.46 | -0.72 | <0.0001 |
| *FSL-FIRST* | 1204.37 | 531.99 | -0.62 | <0.0001 |
| *volBrain* | 1482.54 | 440.99 | -0.66 | <0.0001 |
| *FreeSurfer* | 1902.38 | 412.09 | -0.57 | <0.0001 |
| **Left Hippocampus** | | | | |
| *Stereology* | 187.56 | 412.22 | -0.72 | <0.0001 |
| *FSL-FIRST* | 1160.71 | 527.54 | -0.67 | <0.0001 |
| *volBrain* | 1440.31 | 440.07 | -0.65 | <0.0001 |
| *FreeSurfer* | 1877.69 | 437.74 | -0.57 | <0.0001 |

**Table** 2 *Bias- the mean difference between the manually traced and the volumes from the other techniques. Pearson’s Correlation (r) was computed between the mean differences and their corresponding manual volumes*

|  |  |  |
| --- | --- | --- |
| **Technique** | **Advantages** | **Disadvantages** |
| ***ITK-SNAP*** | * Provides reliable estimates of brain regions/structures. * Flexibility at defining anatomical boundaries prior to segmentation. * Produces segmentation masks for other brain imaging analysis. E.g. these can be co-registered unto fractional anisotropy/mean diffusivity maps for DTI analysis. | * Require investigator expertise of anatomy. * Thorough training required before analysis using intra and inter-rater studies. * High time consumption * High labour intensity * Limited morphometric measures e.g. volume. |
| ***Stereology*** | * Low bias and time efficient means of estimating brain volumes. * Permits the prediction of coefficient of error which provides information to infer the precision of the volumetric extraction. * Permits assessment of regional profiles of structures by virtue of the individual sectioning along slides. * Flexibility at defining anatomical boundaries prior to segmentation. | * Require investigator expertise of anatomy. * Manual intervention required for point counting of the voxels. * Pilot is usually essential to define an exact sampling design to optimise workload for a particular project. * Precision prediction is complex due to the spatial dependence of the observations. * Limited morphometric measures e.g. volume. * Cannot perform shape analysis to investigate localised shape differences. * Does not provide segmentation masks for other brain analysis tasks. |
| ***FSL-FIRST*** | * No need for anatomical expertise. * Freely available for download. * Relatively fast and time efficient (~25mins). * Allows vertex/shape [mesh (vtk)] analysis of subcortical structures to investigate localised shape differences from segmentation masks. * Statistical maps and analysis e.g. multiple comparison correction and other design matrices. * FSLView–visualisation and data management tool | * Relatively fewer segmented output of the brain regions. * Limited options to manually/automatically fix improperly segmented structures. * Output maps/files are in native anatomical space only. * No flexibility at defining subcortical boundaries. |
| ***volBrain*** | * Faster and time efficient (5-12mins) * Freely available for online usewithout requiring a local computer for installation of software. * Produces subcortical segmentation masks/maps in both MNI and native anatomical space. * Automatic asymmetric indices output. * Provides additional neuroanatomical information such as regional macroscopic (hemisphere, cerebellum and brainstem) and intracranial tissue (CSF, WM, GM) volumes. * Relatively lower segmentation failure rate (0.4%) recorded from this study. | * Limited number (n=10) of cases to be processed daily per user. * Cannot perform statistical analysis. * Need other visualisation tools e.g. ITK-SNAP to visualise output data. * Cannot perform shape analysis to investigate localised shape differences. * No flexibility at defining subcortical boundaries. * Minimal anatomical descriptions available defining boundaries employed except the hippocampus (EADC-ADNI Harmonised protocol). |
| ***FreeSurfer*** | * Multiple measures: (volume, area, thickness etc.) of both cortical and subcortical structures. * Frequent software maintenance (upgrades), information and support via a dedicated wikipage plus a quick email response support system. * Freely available for download. * Statistical maps and analysis e.g. cohort comparisons between brains/hemispheres. * Freeview and QDEC-visualisation, analysis and data management tools. * Reconstructed information of cortical surfaces can be obtained based on the segmentation of GM and WM. * Subject-specific parcellation algorithm. * In the field of brain network analyses, it can be used to define brain structures (i.e. nodes) to increase the anatomical sensitivity of findings compared to anatomical templates. It values inter-subject anatomical variability by producing cortico-subcortical measures that are in the individual's own coordinate space, and therefore subject-specific. | * Relatively longer processing time. * High computational power requirement. * Output maps/files are in native anatomical space. * Needs additional toolboxes e.g. LDDMM to perform vertex analysis [as previously described on the semi-automated shape analysis (SASHA) project] and ENIGMA-Shape Pipeline (ENIGMA Shape Analysis). |

**Table 3** *Advantages and disadvantages of the techniques at segmenting subcortical structures*