NeuroImage: Clinical 28 (2020) 102368

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Validity of automated FreeSurfer segmentation compared to manual tracing in detecting prenatal alcohol exposure-related subcortical and corpus callosal alterations in 9- to 11-year-old children

Stevie C. Biffen[a], Christopher M.R. Warton[a], Neil C. Dodge[b], Christopher D. Molteno[c], Joseph L. Jacobson[a,b,c], Sandra W. Jacobson[a,b,c], Ernesta M. Meintjes[d,e,f,*]

[a] Department of Human Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
[b] Department of Psychiatry and Behavioral Neurosciences, Wayne State University School of Medicine, Detroit, MI, USA
[c] Department of Psychiatry and Mental Health, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
[d] Biomedical Engineering Research Centre, Division of Biomedical Engineering, Department of Human Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
[e] Neurosciences Institute, University of Cape Town, South Africa
[f] Cape Universities Body Imaging Centre, University of Cape Town, South Africa

ARTICLE INFO

Keywords:
Prenatal alcohol exposure
Fetal alcohol syndrome
Manual tracing
FreeSurfer
Subcortical volumes
Corpus callosum
Pre-adolescent children

ABSTRACT

In recent years a number of semi-automated and automated segmentation tools and brain atlases have been developed to facilitate morphometric analyses of large MRI datasets. These tools are much faster than manual tracing and demonstrate excellent test–retest reliabilities. Reliabilities of automated segmentations relative to "gold standard" manual tracings have, however, been shown to vary by brain region and in different cohorts. It remains uncertain to what extent smaller brain volumes and potential changes in grey/white matter contrasts in paediatric brains impact on the performance of automated methods, and how pathology may influence performance. This study examined whether using data from automated FreeSurfer segmentation would alter our ability, compared to manual segmentation, to detect prenatal alcohol exposure (PAE)-related volume changes in subcortical regions and the corpus callosum (CC) in pre-adolescent children. High-resolution T1-weighted images were acquired, using a sequence optimized for morphometric neuroanatomical analysis, on a Siemens 3T Allegra MRI scanner in 71 right-handed, 9- to 11-year-old children (27 fetal alcohol syndrome (FAS) and partial FAS (PFAS), 25 non-syndromal heavily exposed (HE) and 19 non-exposed controls) from a high-risk community in Cape Town, South Africa. Data from timeline follow-back interviews administered to the mothers prospectively during pregnancy were used to quantify the amount of alcohol (in ounces absolute alcohol per day, AA/day) that the children had been exposed to prenatally. Volumes of corpus callosum (CC) and bilateral caudate nuclei, hippocampi and nucleus accumbens (NA) were obtained by manual tracing and automated segmentation using both FreeSurfer versions 5.1 and 6.0. Reliability across methods was assessed using intraclass correlation (ICC) estimates for consistency and absolute agreement, and Cronbach's α. Ability to detect regions showing PAE effects was assessed separately for each segmentation method using ANOVA and linear regression of regional volumes with AA/day. Our results support findings from other studies showing excellent reliability across methods for easy-to-segment structures, such as the CC and caudate nucleus. Volumes from FreeSurfer 6.0 were smaller than those from version 5.1 in all regions except the right caudate, for which they were similar, and right hippocampus and CC, for which they were larger. Despite poor absolute agreement between methods in the NA and hippocampus, all three segmentation methods detected dose-dependent volume reductions in regions for which reliabilities on ICC consistency across methods reached at least 0.70, namely the CC, and bilateral caudate nuclei and hippocampi. PAE-related changes in the NA for which ICC consistency did not reach this minimum were inconsistent across methods and should be interpreted with caution. This is the first study to demonstrate in a pre-adolescent cohort the ability of automated segmentation with FreeSurfer to detect regional volume changes associated with pathology similar to those found using manual tracing.

* Corresponding author at: Division of Biomedical Engineering, Department of Human Biology, Faculty of Health Sciences, University of Cape Town, Observatory, 7925, South Africa.
E-mail address: ernesta.meintjes@uct.ac.za (E.M. Meintjes).

https://doi.org/10.1016/j.nicl.2020.102368
Received 1 May 2020; Received in revised form 7 July 2020; Accepted 29 July 2020
Available online 31 July 2020

1. Introduction

As part of ongoing studies of fetal alcohol spectrum disorders (FASD), we recently reported that heavy prenatal alcohol exposure (PAE) is associated with volume reductions in 9- to 11-year-old children in the corpus callosum (CC) and bilaterally in the caudate nucleus and hippocampus using manual tracing (Biffen et al., 2018). Although manual segmentation by hand tracing remains the "gold standard" for measuring structure volumes (Barnes et al., 2008; Boccardi et al., 2011; Dewey et al., 2010; Fischl et al., 2002; Morey et al., 2009) on magnetic resonance (MR) images, it is time consuming and considered impractical for large datasets. Hand tracing of a single complex subcortical structure, such as the hippocampus, can take from 1 to 2 h (Akudjedu et al., 2018; Dewey et al., 2010; Morey et al., 2009) and an entire brain can take 1 week or more to trace (Fischl et al., 2002). Moreover, it has been shown that even when using a single tracer, drift occurs over time due to fatigue (Zou et al., 2004).

In recent years a number of semi-automated and automated segmentation tools and brain atlases have been developed (see Helms, 2016 for a review) to facilitate morphometric analyses of large datasets. Automated segmentation of a single brain performed with the same software on the same hardware and operating system has very high test–retest reliability (Gronenschild et al., 2012). However, when comparing volumes from manual and automated segmentation in both healthy (Cherbuin et al., 2009; Morey et al., 2009; Patenaude et al., 2011) and pathological (Akhondi-Asl et al., 2011; Dewey et al., 2010; Doring et al., 2010; Guenette et al., 2018; Lehmann et al., 2010; Nugent et al., 2013; Pardoe et al., 2009; Pipitone et al., 2014; Sánchez-Benavides et al., 2010; Shen et al., 2010; Tae et al., 2008) adult populations, findings have varied. A recent study in which a typical neuropsychiatric magnetic resonance imaging (MRI) dataset was segmented using three automated segmentation methods (FreeSurfer; FSL-FIRST – FMRIB Integrated Registration and Segmentation Tool, Oxford University, Oxford, UK; volBrain – http://volbrain.upv.es) demonstrated good partial correlation (0.77–0.87) to manually traced volumes on an "easy-to-segment" structure, such as the caudate nucleus, while a more difficult structure, such as the hippocampus, achieved moderate correlation (0.35–0.62) but poor absolute agreement (intraclass correlation (ICC) 0.07–0.10) (Akudjedu et al., 2018). While absolute volume differences may reflect differences in segmentation protocol, correlation analyses provide a measure of consistency between methods. Taken together, findings from previous studies suggest that performance of automated methods depends on the segmented structure as well as the protocol used and that discrepancy is greater in atrophic brains (Sánchez-Benavides et al., 2010) and smaller brain regions (Guenette et al., 2018; Lehmann et al., 2010; Schoemaker et al., 2016).

To date, it remains uncertain to what extent smaller brain volumes and potential changes in grey/white matter contrasts in paediatric brains impact on the performance of automated methods. We found only one study in children (age 6–11 yr) that examined accuracy of automated subcortical segmentation (Schoemaker et al., 2016). Hippocampal and amygdala volumes using FreeSurfer and FSL-FIRST, respectively, showed moderate (0.61–0.77) and weak (0.31–0.59) correlations with those from manual segmentation, and all ICCs, except left hippocampus volume from FreeSurfer, failed to reach 0.70 (Schoemaker et al., 2016). Performance may be worse in paediatric conditions, such as fetal alcohol syndrome (FAS), that are characterised by small head circumference and reduced total brain volume (Hoyme et al., 2005).

Since the aim of many research studies is to examine brain changes associated with pathology, the ability and accuracy of automated techniques in distinguishing individuals in a clinical group from healthy controls may be more important than absolute agreement. While FreeSurfer has been shown to perform reasonably well in this regard in patients with Alzheimer's Disease, demonstrating volume reductions (Lehmann et al., 2010; Shen et al., 2010) and hippocampal atrophy rates (Mulder et al., 2014) similar to manual segmentation, automated methods have been less successful in distinguishing between groups or identifying associations with behavioural/clinical outcomes in other pathologies. For example, in patients with HIV, the association of caudate, putamen, amygdala, and hippocampal volumes with clinical measures of disease progression differed for outputs generated by FreeSurfer, IBASPM (Individual Brain Atlases using Statistical Parametric Mapping) and auto-assisted manual tracings (Dewey et al., 2010). Depression-related hippocampal volume reductions were detected with FreeSurfer but not FSL-FIRST (Morey et al., 2009), and in former National Football League (NFL) players with neurobehavioral symptoms, automated FreeSurfer segmentation identified group differences relative to age-matched controls in 4 of 11 regions, compared to 8 of 11 with manual correction, as well as different regions showing associations with neurobehavioral factors (Guenette et al., 2018). Due to an absence of subgroups, Schoemaker et al. (2016) could not examine this question in their study on segmentation accuracy in children. Notably, in our research on 5-year-old children with HIV, automated segmentation with FreeSurfer yielded no volumetric differences compared to uninfected controls in basal ganglia, other than reductions in left globus pallidus and a tendency to larger corpus callosi. In contrast, manual segmentation demonstrated HIV-related volume *increases* in the left globus pallidus and bilaterally in the nucleus accumbens (NA) and putamen, as well as CC *reductions* (Randall et al., 2017).

In the present work, we examine whether using data from an automated FreeSurfer segmentation would alter our ability compared to manual segmentation to detect PAE-related volume changes in subcortical regions and the corpus callosum (CC) in school-aged children. To date, the impact of the segmentation method when assessing effects of PAE has only been investigated in the cerebellum (Cardenas et al., 2014) using the Cerebellar Analysis Toolkit (CATK), which was able to detect changes related to PAE in 20 children and adolescents (aged 10–18 yr) similar to those found using manual tracing. Despite a known tendency to overestimate subcortical volumes (Cherbuin et al., 2009; Dewey et al., 2010; Doring et al., 2010; Schoemaker et al., 2016; Tae et al., 2008), we chose to use FreeSurfer, as it has been shown to perform better than other automated methods on some of the structures we were investigating (Dewey et al., 2010; Morey et al., 2009; Schoemaker et al., 2016) with greater power to detect group differences (Morey et al., 2009). Since FreeSurfer was updated during our study, we present comparisons to outputs from manual tracing for two versions to examine progress towards gold standard manual segmentation. We hypothesized that data from both manual tracing and FreeSurfer would demonstrate PAE-related volumetric reductions in the CC and bilaterally in the caudate nuclei and hippocampi.

2. Methods

2.1. Participants

Participants consisted of children from our Cape Town Longitudinal Cohort, who were born to Cape Coloured (mixed ancestry) pregnant women recruited from 1998 to 2002 from an antenatal clinic in a historically disadvantaged community in Cape Town, South Africa (Jacobson et al., 2008), where prevalence of alcohol abuse is unusually high and fetal alcohol syndrome (FAS) is among the highest in the world (May et al., 2013). Pregnant women were screened and recruited at clinic enrolment and interviewed twice during pregnancy and once within 1 month postpartum (to capture 3rd trimester drinking) regarding alcohol consumption using the timeline follow-back approach (Jacobson et al., 2002), which includes questions about how much alcohol (beer, wine, hard liquor) the mother consumed over a 2-week period preceding the interview. The volume of alcohol consumed was then converted to ounces of absolute alcohol (oz AA; 1 oz AA is the equivalent of ≈2 standard drinks) and three measures of alcohol consumption computed – oz AA per day (AA/day), oz AA per drinking

occasion (AA/drinking day), and frequency of drinking (drinking days/week). At time of recruitment binge drinking was 5 or more drinks/occasion. Any woman who reported alcohol consumption of at least 1 oz AA/day or at least 2 instances of binge drinking within her first trimester was invited to participate in the study. Pregnant women from the same clinic were invited to participate as controls if they reported abstaining or minimal alcohol consumption and no binge drinking. Participants were also interviewed at each of these visits about their illicit drug use (marijuana, methaqualone ("mandrax") and cocaine use/day) and number of cigarettes smoked per day during pregnancy.

Exclusion criteria for participation were age < 18 years or chronic medical problems, such as diabetes, epilepsy or cardiac problems. Infants from multiple births were excluded, as well as those presenting with major chromosomal anomalies, seizures and neural tube defects.

The children were independently examined by two expert dysmorphologists (H.E. Hoyme, M.D. (H.E.H.) and L.K. Robinson, M.D.) at FASD diagnostic clinics we organized in 2005 and 2009. Children were examined for growth and FAS-related dysmorphic features using a standard protocol based on the Revised Institute of Medicine criteria (Hoyme et al., 2005). The determination of which children met criteria for diagnosis with FAS or partial FAS (PFAS) was made during case conferences with the dysmorphologists and SWJ, JLJ, CDM. If children did not fall into either of these diagnostic groups they were placed, based on maternal alcohol consumption, into either the non-syndromal heavily exposed (HE) group or the control group. Diagnoses of the children were confirmed by examinations in follow-up diagnostic clinics in 2013 and 2016, in which H.E.H. was again the lead dysmorphologist. For purposes of the present analyses, data from the two syndromal groups, FAS and PFAS, were combined.

2.2. Magnetic resonance image (MRI) acquisition

81 right-handed children were scanned at the Cape Universities Brain Imaging Centre (CUBIC) using a 3T Siemens Allegra MRI (Siemens Medical Systems, Erlangen, Germany) when they were 9–11 years old (Meintjes et al., 2014). High-resolution T1-weighted images were obtained using a volumetric navigated (Tisdall et al., 2012) multi-echo magnetization prepared rapid gradient echo (MEMPRAGE) sequence. This sequence was optimized for analysis with FreeSurfer (van der Kouwe et al., 2008). Imaging parameters were: FOV 256 × 256 mm^2, 128 sagittal slices, TR 2530 ms, TE 1.53/3.21/4.89/6.57 ms, TI 1100 ms, flip angle 7°, voxel size 1.3 × 1.0 × 1.3 mm^3, acquisition time 8:07 min. Real-time motion tracking and correction by the volumetric navigator reduced artefacts resulting from motion.

Ethics approval was obtained from the human research ethics committees of the University of Cape Town Faculty of Health Sciences and Wayne State University. Written informed consent was obtained from the mothers/guardians and oral assent from the children.

2.3. Segmentation protocol

A blinded neuroanatomical researcher (S.B.) traced the NA, caudate nuclei, hippocampi and CC manually on MR images with Multitracer (Woods, 2003) software on a Wacom tablet laptop (Lenovo ThinkPad X200) using a previously described protocol (Biffen et al., 2018). Briefly, hippocampal tracings only included grey matter and neither the alveous (a white matter sheath surrounding the anterior hippocampus) nor the fornices (white matter projections from the hippocampus). The caudate nucleus and NA were traced together using the lateral ventricle as the medial border and surrounding white matter (WM) as the lateral border. These two structures were separated by drawing a straight line between the inferior border of the lateral ventricle and the border of the internal capsule. Multitracer "frust" volumes were computed for these structures. The CC was traced on two contiguous midline sagittal slices and averaged to give CC area. Images of 10 randomly selected participants were re-traced by S.B. to assess intra-rater reliabilities, and by another blinded neuroanatomical researcher (S.R.) to assess inter-rater reliabilities.

Automated segmentation and parcellation were performed using both FreeSurfer versions 6.0.0 (FS6.0) and 5.1.0 (FS5.1). Segmentations were visually checked for accuracy; no manual corrections were required.

Ten participants were excluded – three due to pathology unrelated to the research question (2 HE; 1 control) and seven due to compromised image quality (1 FAS; 2 HE; 4 controls). With regards the three pathological exclusions: two (1 HE, 1 control) were due to the inferior boundary of the frontal horn of the lateral ventricle not being visible, which meant that the delineation separating the NA and caudate nucleus could not be performed; one HE participant was excluded due to abnormal hypointensity superior to the lateral ventricle.

2.4. Statistical analyses

Statistical analyses were performed in SPSS statistical package version 25 (SPSS Inc, Chicago, IL).

2.4.1. Reliability

Test-retest and inter-rater reliabilities of manual tracings were assessed using Pearson correlation, as well as ICC based on a mean-rating two-way mixed-effects model. We report ICC estimates for both consistency (correspondence) and absolute agreement (presence of systematic differences). The convention proposed by Koo and Li (2016) was adopted whereby ICC estimates < 0.5 indicate poor reliability, values between 0.5 and 0.75 moderate reliability, between 0.75 and 0.9 good reliability, and values ≥0.9 excellent reliability.

ICC estimates for consistency and absolute agreement based on a single-rating two-way mixed-effects model, as well as Cronbach's α and Bland-Altman plots, were used to test agreement between regional volumes obtained from the three different segmentation methods. In view of the fact that FreeSurfer computes a CC volume and our manual tracing protocol an area on a mid-sagittal slice, normalised z-scores were used in the analyses involving the CC volumes/areas.

2.4.2. Validity

To examine whether the three segmentation methods performed similarly across diagnostic groups, we performed a repeated measures general linear model (GLM) for each region with main effects of group and method, and group by method interactions.

Regions showing volume differences between diagnostic groups were additionally identified for each segmentation method separately using ANOVAs. Potential confounders, including child sex and age at scan (yr), socioeconomic status (SES; Hollingshead, 2011), maternal smoking during pregnancy (cigarettes/day), and postnatal lead exposure (μg/dl) obtained from a blood sample at 5 years, were subsequently controlled for in the group comparisons using ANCOVA. An additional ANCOVA was run to examine whether group differences persisted after adjustment for TIV. It should be noted that linearly controlling for total brain volume in this way neglects the allometric phenomenon whereby different brain regions may exhibit different scaling relations to TIV. For manually traced data, TIVs from FS6.0 were used.

Since degree of alcohol exposure often demonstrates greater sensitivity to detect PAE-related pathology than FASD diagnosis, we additionally examined the association between a continuous measure of alcohol exposure (oz AA/day) and structure volumes (or CC area for manually traced data) using Pearson correlation for each segmentation method separately. Multiple regression was used to control for potential confounders, and TIV was included in an additional step to determine whether regions were disproportionately affected compared to whole brain volume reductions typically seen in PAE.

Finally, multiple regression analyses were used to compare differences in effect size generated by each of the segmentation methods.

Separate regressions were run for each method. PAE functioned as the "dependent variable" in these analyses; confounders without and with TIV were entered in Steps 1 and 2, respectively; the volumes of the three most affected regions, in Step 3. The additional variance in PAE explained by the volumes jointly at Step 3 provides a summary measure of the effect size associated with each segmentation method.

3. Results

Sample characteristics for the 71 children and their mothers are summarized in Table 1. Children in the FAS/PFAS and control groups were slightly younger than those in the HE group. Those in the FAS/PFAS group had lower IQs, and TIVs were smaller for the syndromal FAS/PFAS group than the non-syndromal HE and control groups. Groups did not differ by sex or childhood lead exposure. There was a tendency for mothers of children in the FAS/PFAS group to smoke more cigarettes than those in the control group.

By design, mothers of children in the FAS/PFAS and HE groups consumed more alcohol than mothers of control children, with FAS/PFAS mothers consuming more than the HE mothers except for AA/drinking day. The key difference between mothers of the FAS/PFAS and HE groups was that the FAS/PFAS mothers binge drank more frequently – drinking twice as often per week. All but one mother in the control group abstained from drinking during pregnancy – that mother consumed only 2 drinks/occasion on 2–3 days during her pregnancy. Marijuana and cocaine use were rare and did not differ by group; none of the mothers reported using methaqualone.

3.1. Reliability

In all brain regions traced manually, except left NA, test–retest reliabilities were greater than 0.93 on all tests performed, indicating that the same rater can repeatedly trace the same structures with excellent agreement. In left NA, values on the different tests ranged from 0.74 (Pearson correlation) to 0.84 (ICC absolute agreement).

Inter-rater ICCs for consistency and absolute agreement were ≥ 0.95 in all regions, except right hippocampus and bilateral NA. In right hippocampus and right NA, ICCs were between 0.90 and 0.95; and in left NA, both consistency and agreement were worst at 0.83. Inter-rater Pearson correlations were ≥ 0.90 in all regions except right hippocampus and left NA, where values were 0.89 and 0.71, respectively. Overall, these results suggest that another trained neuroanatomist would obtain similar results on the same ROIs given the same manual tracing protocol, even in the left NA, where reliabilities would at a minimum be acceptable.

When comparing volumes from the three segmentation methods, the caudate nuclei and z-normalised CC demonstrated good to excellent consistency and absolute agreement (Table 2). Overall, ICC consistency across all three methods for NA and hippocampi were moderate (≥ 0.5 but < 0.75) but absolute agreement poor (< 0.5). In hippocampi, ICC consistency across all three methods at least reached 0.70, which has often been considered a minimum standard for adequate reliability (Nunnally et al., 1967; Terwee et al., 2007). Notably, consistency in NA between manual tracing and FS5.1 failed to reach 0.5, while consistency between manual tracing and FS6.0 fell just short of reaching 0.5 for the left NA and was greater on the right, pointing to improvements in automated segmentation with version updates. The fact that

Table 1
Sample characteristics.

		Exposed								
		Control		HE		FAS/PFAS		Total		F or X^2
		(n = 19)		(n = 25)		(n = 27)		(N = 71)		
Child:										
Sex: Male	n(%)	9	(47%)	16	(64%)	14	(52%)	39	(55%)	1.37
Age at scan (yr)[a]	Mean (SD)	10.6	(0.5)	11.0	(0.7)	10.5	(0.6)	10.7	(0.7)	4.05*
WISC-IV IQ[b]	Mean (SD)	73	(10.7)	76	(11.1)	64	(11.0)	71	(12.1)	8.73**
TIV ($\times 10^6$ mm^3)[c]	Mean (SD)	1.407	(0.085)	1.457	(0.134)	1.319	(0.166)	1.391	(0.015)	6.76*
Lead (ug/dl)	Mean (SD)	9.2	(3.5)	9.4	(3.3)	11.6	(5.9)	10.2	(4.6)	2.16
Maternal:										
Cigarettes/day[d,#]	Median(IQR)	0.0	(5.0)	5.3	(8.3)	7.5	(5.0)	5.0	(10.0)	2.38[†]
Marijuana (yes)	n (%)	0	(0%)	2	(8%)	3	(11%)	5	(7%)	2.16
Cocaine (yes)	n (%)	0	(0%)	0	(0%)	1	(3.7%)	1	(1%)	1.65
Alcohol at conception										
AA/day (oz)[e]	Median(IQR)	0.0	(0.0)	0.6	(1.0)	1.3	(1.3)	0.6	(1.3)	35.67***
AA/drinking day (oz)[f]	Mean (SD)	0.06	(0.3)	2.8	(2.2)	4.6	(2.4)	2.8	(2.7)	28.50***
Drinking days/wk[g]	Mean (SD)	0.0	(0.1)	1.4	(1.1)	2.7	(1.6)	1.5	(1.6)	26.90***
Alcohol across pregnancy										
AA/day (oz)[h]	Median(IQR)	0.0	(0.0)	0.2	(0.7)	1.0	(0.8)	0.3	(1.0)	31.61***
AA/drinking day (oz)[i]	Mean(SD)	0.06	(0.3)	3.4	(2.4)	4.2	(1.8)	2.8	(2.5)	31.04***
Drinking days/wk[j]	Mean (SD)	0.0	(0.0)	1.0	(0.8)	2.0	(1.3)	1.1	(1.2)	26.47***

For skewed data medians and interquartile ranges (IQR) were used. HE heavily exposed non-syndromal; FAS fetal alcohol syndrome; PFAS partial FAS; WISC-IV Wechsler Intelligence Scales for Children-Fourth Edition; AA absolute alcohol; TIV total intracranial volume as measured by FreeSurfer v. 6.0.

[#] Winsorized data used (PFAS/FAS n = 0, HE n = 0, Control n = 1).

[a] HE > FAS/PFAS ($p = 0.007$), Control ($p = 0.06$).

[b] FAS/PFAS < HE ($p < 0.001$), Control ($p = 0.006$).

[c] FAS/PFAS < HE ($p = 0.001$), Control ($p = 0.035$).

[d] FAS/PFAS > Control ($p = 0.035$).

[e] Control < FAS/PFAS, HE (both $p < 0.001$); HE < FAS/PFAS ($p < 0.001$).

[f] Control < FAS/PFAS, HE (both $p < 0.001$); HE < FAS/PFAS ($p = 0.002$).

[g] Control < FAS/PFAS, HE (both $p < 0.001$); HE < FAS/PFAS ($p < 0.001$).

[h] Control < FAS/PFAS, HE (both $p < 0.001$); HE < FAS/PFAS ($p < 0.001$).

[i] Control < FAS/PFAS, HE (both $p < 0.001$).

[j] Control < FAS/PFAS, HE (both $p \leq 0.001$); HE < FAS/PFAS ($p < 0.001$).

[†] $p \leq 0.1$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Table 2
Reliabilities of regional volumes obtained from manual tracing and automated segmentation with FreeSurfer versions 5.1 and 6.0, respectively, as reflected by intraclass correlation (ICC) coefficients and 95% confidence intervals (CIs) for consistency and absolute agreement, and Cronbach's α (N = 71).

	PAIRWISE COMPARISONS						COMPARISON OF ALL THREE SEGMENTATION METHODS		
	FS5.1 vs MANUAL		FS6.0 vs MANUAL		FS5.1 vs FS6.0				
Regions	Consistency (95% CI)	Absolute Agreement (95% CI)	Consistency (95% CI)	Absolute Agreement (95% CI)	Consistency (95% CI)	Absolute Agreement (95% CI)	Consistency (95% CI)	Absolute Agreement (95% CI)	Cronbach's α
TIV[1]	–	–	–	–	0.69 (0.55–0.80)	0.63 (0.36–0.78)	–	–	0.82
Grey matter ROIs									
L Caudate	0.86 (0.79–0.91)	0.86 (0.78–0.91)	0.89 (0.83–0.93)	0.87 (0.75–0.93)	0.93 (0.88–0.95)	0.88 (0.57–0.95)	0.89 (0.84–0.93)	0.87 (0.79–0.92)	0.96
R Caudate	0.86 (0.78–0.91)	0.85 (0.77–0.91)	0.91 (0.86–0.94)	0.91 (0.86–0.94)	0.95 (0.93–0.97)	0.94 (0.89–0.97)	0.91 (0.86–0.94)	0.90 (0.86–0.93)	0.97
L NA	0.34 (0.11–0.53)	0.13 (−0.08–0.37)	0.49 (0.29–0.65)	0.22 (−0.09–0.52)	0.67 (0.52–0.78)	0.60 (0.31–0.77)	0.50 (0.36–0.63)	0.26 (0.01–0.49)	0.75
R NA	0.40 (0.19–0.58)	0.08 (−0.05–0.28)	0.63 (0.47–0.75)	0.25 (−0.08–0.59)	0.70 (0.55–0.80)	0.39 (−0.10–0.72)	0.57 (0.45–0.69)	0.19 (−0.02–0.43)	0.80
L Hippocampus	0.60 (0.43–0.73)	0.06 (−0.02–0.24)	0.67 (0.51–0.78)	0.08 (−0.02–0.31)	0.83 (0.74–0.89)	0.74 (0.32–0.88)	0.70 (0.59–0.79)	0.11 (−0.01–0.31)	0.87
R Hippocampus	0.68 (0.53–0.79)	0.08 (−0.02–0.30)	0.70 (0.56–0.80)	0.07 (−0.02–0.29)	0.85 (0.77–0.90)	0.83 (0.71–0.90)	0.75 (0.65–0.82)	0.12 (−0.01–0.33)	0.90
White matter ROI									
CC[2]	0.85 (0.77–0.90)	0.85 (0.77–0.90)	0.84 (0.75–0.90)	0.84 (0.75–0.90)	0.95 (0.93–0.97)	0.95 (0.93–0.97)	0.88 (0.83–0.92)	0.88 (0.83–0.92)	0.96

TIV Total intracranial volume; NA Nucleus accumbens; CC corpus callosum; L left; R right; ICC estimates based on a single-rating two-way mixed-effects model.
[1] TIV data only available for automated segmentations using FreeSurfer.
[2] Analyses used z-normalised CC volumes (for automated segmentation) and areas (for manual segmentation).

consistency between FS5.1 and FS6.0 in NA was only moderate, and agreement poor to moderate, highlight that automated segmentation of this region may not be reliable. Internal consistency, reflected by Cronbach's α, was ≥0.87 in all regions except the NA. Although TIV estimates from versions 5.1 and 6.0 showed moderate consistency and agreement, those from version 6.0 were significantly higher than those from version 5.1 (FS5.1 mean ± sd: (1.326 ± 0.014) × 10^6 mm^3; FS6.0: (1.391 ± 0.015) × 10^6 mm^3; $p < 0.001$).

Visual inspection of Bland-Altman plots of agreement between FS6.0 and manual tracing (Fig. 1, left panel), and the two versions of FreeSurfer (Fig. 1, right panel), demonstrated no associations of volume differences between methods and average volumes in any region, but confirmed substantial bias in NA and hippocampi where manually traced volumes were smaller than those from FS6.0.

3.2. Validity

Fig. 2 shows box-and-whisker plots comparing subcortical volumes from the three segmentation methods separately within each diagnostic group, together with differences (at $p < 0.05$) on *posthoc* pairwise comparisons. Repeated measures GLMs revealed main effects of segmentation method in all regions (all p's ≤ 0.008) except CC where z-normalised volumes/areas showed no differences between methods ($p = 0.99$), and main effects of group in bilateral NA and hippocampi (all p's ≤ 0.005), and below conventional levels of significance in left caudate ($p = 0.06$) and CC ($p = 0.08$). Interactions between method and diagnostic group were evident in the right NA ($p = 0.001$) and left hippocampus ($p = 0.02$), and below conventional levels of significance in the right caudate ($p = 0.07$) and right hippocampus ($p = 0.08$). In right NA and left hippocampus, the method by group interaction was attributable to FS5.1 and FS6.0 demonstrating smaller volumes in children with FAS/PFAS, not seen with manual tracing (see Table 3). In right caudate, volume differences between segmentation methods seen in HE children were not apparent in the other groups (Fig. 2), and in right hippocampus, FS5.1 volumes were smaller than those from FS6.0 only in HE children. Despite these interaction effects, Fig. 2 shows that the pattern in which volumes differed between the three segmentation methods were generally similar across diagnostic groups. Notably, except in the caudate nuclei, manually traced volumes were typically smaller than those from either version of FreeSurfer, and volumes from FS6.0 were generally smaller than those from FS5.1. Exceptions to the latter being right caudate and right hippocampus in control and FAS/PFAS groups where FS5.1 and FS6.0 volumes were similar, and right hippocampus in HE children and CC where volumes from FS6.0 were larger than those from FS5.1.

When assessing agreement of methods in detecting effects of PAE on regional brain volumes, manually traced data only demonstrated reductions in right hippocampus in children with FAS/PFAS compared to HE and control children (Table 3A), while automated segmentation demonstrated reductions in FAS/PFAS children compared to both other groups (all *posthoc* p's ≤ 0.05) unilaterally in caudate (right on FS5.1; left on FS6.0) and bilaterally in NA and hippocampus (Table 3B and C). Virtually all of the findings remained significant after control for confounders (column p_1 in Table 3). When adjusting for TIV (column p_2 in Table 3), only right hippocampus tended to be disproportionately smaller on volumes from manual tracing and FreeSurfer v.6.0. By contrast, volumes from the older FreeSurfer v.5.1 demonstrated disproportionate reductions in right NA and bilateral hippocampus in children with FAS/PFAS compared to the other groups.

All three segmentation methods demonstrated association of increasing PAE with decreasing caudal and hippocampal volumes – effects that largely remained significant after control for confounders and TIV (Table 4 and Fig. 3). Except for right NA in FS5.1, associations of increasing PAE with smaller NA volumes seen only in data from automated segmentation, did not survive after adjustment for TIV. Although higher PAE was related to smaller CC size across all methods, the

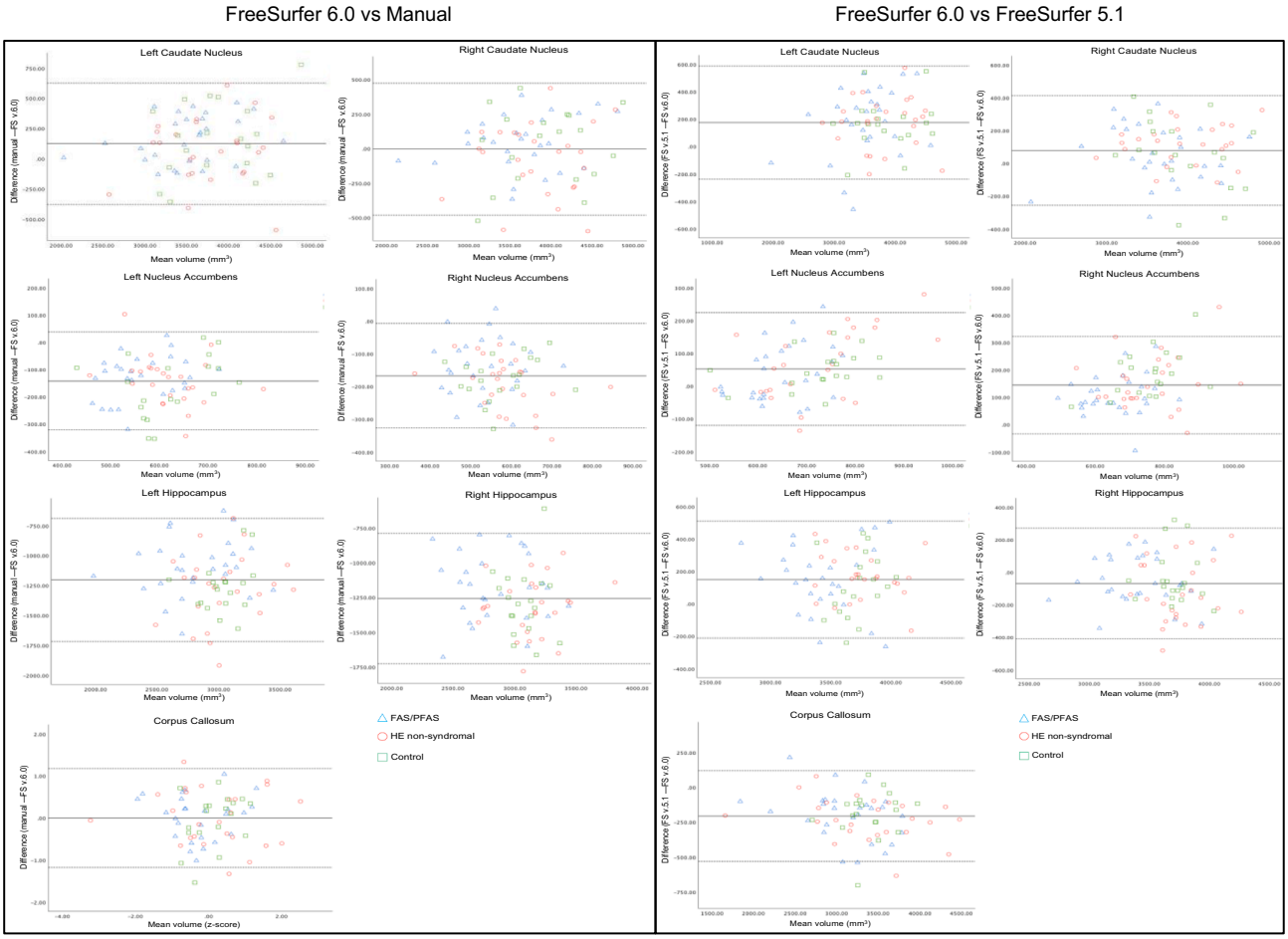

Fig. 1. Bland-Altman plots of agreement between (left) manual tracing and FreeSurfer version 6.0 (FS6.0), and (right) FS5.1 and FS6.0. Comparisons involving manually traced CC areas used z-normalised areas (manual) and volumes (FS6.0). The solid lines represent the mean differences, and the dashed lines, the mean ± 1.96 standard deviations.

association did not survive in manually traced data after control for confounders, nor in any of the methods after adjustment for TIV. Comparing the slopes using a *t*-test (Soper, 2020; Cohen et al., 2003) revealed no differences (all p's ≥ 0.14) between methods in any regions except CC, for which the slope of the manually traced areas as a function of oz AA consumed per day (slope ± standard error = −69 ± 24 mm^2/oz AA/day) was smaller than with CC volumes from FS5.1 (−456 ± 132 mm^3/oz AA/day, p = 0.005) or FS6.0 (−513 ± 141 mm^3/oz AA/day, p = 0.002), and right NA for which the slope with manually traced volumes was smaller than with FS5.1 volumes (manual: −18 ± 25; FS5.1: −113 ± 34, p = 0.03).

Multiple regression analyses of amount of PAE as a function of confounders and TIV, demonstrated significant improvement when hippocampal, caudal and CC volumes from FS6.0 were added to the model, but not when hippocampal and caudal volumes and CC area from manual tracing were added (Table 5). NA volumes were not included in these models due to their lower reliabilities and absence of associations with PAE in manually traced data.

4. Discussion

4.1. Reliability

Our results support findings from other studies showing good to excellent consistency and agreement between manual and automated segmentations for the caudate nucleus and CC, which are relatively easy to segment (Akudjedu et al., 2018). Across FASD diagnostic groups, these regions also demonstrated fewer and smaller volume differences between methods than other regions (Fig. 2). In contrast, for the NA and hippocampus that are more difficult to delineate, consistency was moderate but absolute agreement poor (< 0.26). Schoemaker et al. (2016) reported similar ICC consistencies between manual and automated FreeSurfer (v4.4) segmentation in children aged 6–11 years of 0.74 (CI: 0.66–0.81) and 0.68 (CI: 0.59–0.76) for the left and right hippocampus, respectively, but overestimation of volumes by 60.4% and 51.5%, respectively. The poor absolute agreement scores may reflect differences in segmentation protocols, while the moderate consistency scores suggest somewhat reliable systematic differences. Notably, Cronbach's α values were substantially higher than ICC estimates in all regions. Commonly accepted criteria for Cronbach's α suggest that internal consistency should not be below 0.80 in basic research and should be above 0.90 for widely used scales; α ≥ 0.95 should be the desired standard for any scale (Carmines and Zeller, 1979; Lance et al., 2006). Our findings here of internal consistency across methods ≥0.80 in all regions except the left NA, despite varying results on ICC, highlight why this measure should be interpreted with caution and not in isolation when assessing reliability.

Since it has been reported that discrepancy of automated segmentation is greater in atrophic brains (Sánchez-Benavides et al., 2010), it is encouraging that a comparison of segmentation methods separately within each diagnostic group yielded similar results in children that had been prenatally exposed to alcohol compared to controls. Within each of the diagnostic groups, manually traced NA and hippocampi were smaller than those from automated segmentations using either version

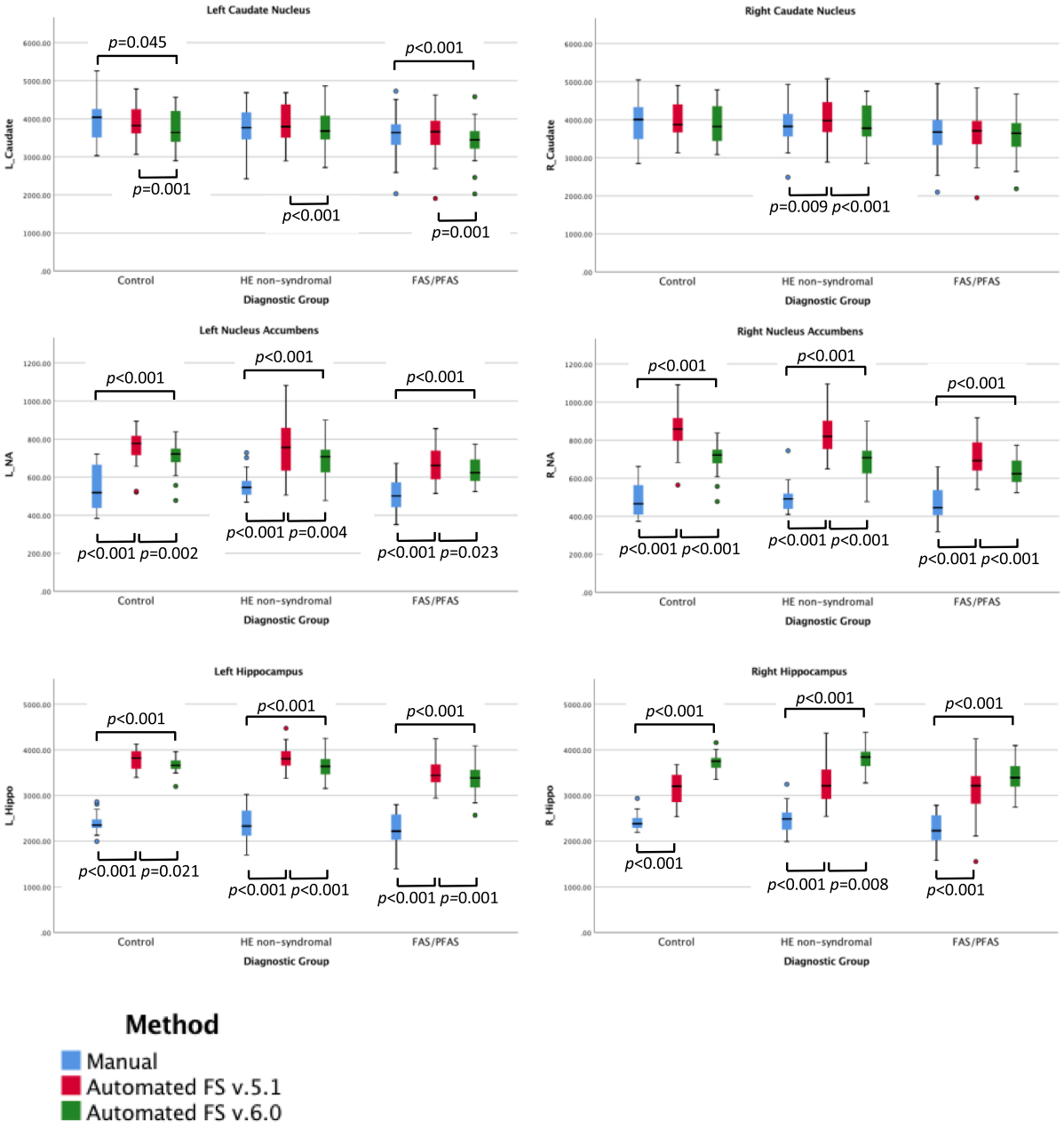

Fig. 2. Box-and-whisker plots comparing subcortical volumes (mm^3) from the three segmentation methods within diagnostic groups. Brackets indicate differences between methods (at $p < 0.05$) on *posthoc* pairwise comparisons.

of FreeSurfer. Schoemaker et al. (2016) similarly found that FreeSurfer, albeit an earlier version, overestimated hippocampal and amygdala volumes in 6- to 11-year-old children. Differences between volumes from the two FreeSurfer versions were also consistent across diagnostic groups. In all regions except right caudate, right hippocampus and CC, volumes from FS6.0 were smaller than those from FS5.1. These differences point to changes and potentially greater accuracy in the delineation of subcortical structures on the newer version of FreeSurfer.

It is interesting that comparisons of automated and manual segmentation in the caudate nuclei and right hippocampus yielded slightly different results in HE children compared to the other two groups. Specifically, right caudal volumes on FS5.1 that were similar to other methods in controls and children with FAS/PFAS were larger in HE children; left caudal volumes from FS6.0 that were smaller than those from manual tracing in other groups were similar in HE children; and finally, right hippocampal volumes on FS6.0 that were similar to those from FS5.1 in controls and children with FAS/PFAS were larger in HE children. While we do not have explanations for these discrepancies, it is possible that greater diversity in the HE group may play a role. These children's mothers often drank as heavily, albeit less frequently, than mothers of children with FAS/PFAS, and it is not known why or to what extent these children may have been spared.

Figs. 4–7 illustrate on comparable slices in the same subjects differences between manual and automated segmentations of the regions examined here, as well as inaccurate inclusion by FreeSurfer of external structures that may contribute to overestimation of volumes. The dotted lines in Fig. 4A indicate where the superior borders of the NA were drawn on the caudate nuclei tracings. The fact that we created a theoretical, rather than organic, border to separate the NA from the caudate nucleus in our manual tracing protocol (Biffen et al., 2018; Looi et al., 2008; Randall et al., 2017) probably contributed to the fact that consistency across methods was worst and absolute agreement poor in

Table 3
Comparison of regional volumes by diagnostic group for each of the three segmentation methods separately.

ROI	Control ($n=19$)		HE ($n=25$)		FAS/PFAS ($n=27$)		Total ($N=71$)		% Difference relative to controls		p	p_1	p_2
									FAS/PFAS	HE			
A. MANUAL TRACING[a]													
Grey matter ROIs	Volumes (mm^3)												
L Caudate	3929	(563)	3785	(536)	3585	(573)	3748	(567)	−8.7	−3.7	0.118	0.155	0.356
R Caudate	3936	(582)	3806	(541)	3655	(617)	3783	(584)	−7.1	−3.3	0.272	0.294	0.406
L NA	542	(117)	549	(71)	501	(90)	529	(93)	−7.6	1.3	0.143	0.051	0.320
R NA	490	(95)	492	(82)	464	(90)	481	(88)	−5.3	0.4	0.472	0.176	0.810
L Hippocampus	2400	(224)	2394	(339)	2272	(365)	2349	(325)	−5.3	−0.3	0.294	0.213	0.722
R Hippocampus	2432	(185)	2473	(296)	2255	(319)	2379	(294)	−7.3	1.7	**0.016**	**0.002**	0.065
White matter ROI	Areas (mm^2)												
CC	564	(64)	569	(105)	524	(80)	551	(88)	−7.1	0.9	0.128	0.452	0.896
B. AUTOMATED SEGMENTATION FreeSurfer v.5.1.0[b]													
Grey matter ROIs	Volumes (mm^3)												
L Caudate	3942	(509)	3897	(525)	3606	(591)	3798	(560)	−8.5	−1.1	0.072	0.090	0.193
R Caudate	3977	(468)	4000	(542)	3662	(562)	3865	(548)	−7.9	0.6	**0.047**	0.070	0.265
L NA	753	(104)	757	(148)	670	(97)	723	(125)	−11.0	0.5	**0.017**	0.108	0.212
R NA	848	(118)	836	(127)	713	(101)	792	(130)	−15.9	−1.4	**< 0.001**	**0.001**	**0.002**
L Hippocampus	3780	(219)	3839	(262)	3520	(325)	3702	(310)	−6.9	1.6	**< 0.001**	**< 0.001**	**0.013**
R Hippocampus	3685	(185)	3681	(261)	3379	(333)	3567	(309)	−8.3	−0.1	**< 0.001**	**0.004**	**0.028**
White matter ROI													
CC	3302	(320)	3265	(624)	3008	(447)	3177	(502)	−8.9	−1.1	0.080	0.234	0.586
C. AUTOMATED SEGMENTATION FreeSurfer v.6.0.0[b]													
Grey matter ROIs	Volumes (mm^3)												
L Caudate	3777	(487)	3723	(524)	3423	(516)	3624	(529)	−9.4	−1.4	**0.039**	**0.037**	0.341
R Caudate	3915	(547)	3882	(527)	3611	(568)	3787	(558)	−7.8	−0.8	0.110	0.123	0.534
L NA	704	(85)	689	(95)	630	(70)	670	(88)	−10.5	−2.1	**0.007**	**0.006**	0.174
R NA	663	(81)	682	(113)	605	(90)	648	(101)	−8.7	2.9	**0.016**	**0.010**	0.508
L Hippocampus	3667	(173)	3667	(291)	3364	(336)	3552	(317)	−8.3	0	**< 0.001**	**0.003**	0.169
R Hippocampus	3723	(192)	3796	(269)	3429	(332)	3637	(321)	−7.9	2.0	**< 0.001**	**< 0.001**	0.059
White matter ROI													
CC	3492	(291)	3493	(672)	3204	(508)	3383	(541)	−8.2	0	0.091	0.325	0.960

Values are Mean (SD); NA nucleus accumbens; CC corpus callosum; TIV total intracranial volume; HE heavily exposed non-syndromal; FAS fetal alcohol syndrome; PFAS partial FAS.
Bold print denotes significance at $p \leq 0.05$.
p_1: p-value after controlling for potential confounders (child sex, age, and lead concentration (ug/dl); maternal smoking during pregnancy and socioeconomic status).
p_2: p-value after controlling for TIV in addition to potential confounders (child sex, age and lead concentration (ug/dl); maternal smoking during pregnancy and socioeconomic status).
[a] FreeSurfer v.6.0 TIV used.
[b] FreeSurfer v.5.1 TIV used.

Table 4
Association of amount of prenatal alcohol exposure with subcortical and corpus callosum sizes for each of the segmentation methods separately.

	Manual Tracing			FreeSurfer v. 5.1			FreeSurfer v. 6.0		
ROI	r (p)	β_1 (p)	β_2[a] (p)	r (p)	β_1 (p)	β_2[b] (p)	r (p)	β_1 (p)	β_2[a] (p)
L Caudate	**−0.366**	**−0.418**	**−0.271**	**−0.335**	**−0.435**	**−0.351**	**−0.353**	**−0.467**	**−0.320**
	(0.002)	**(0.003)**	**(0.035)**	**(0.004)**	**(0.002)**	**(0.007)**	**(0.002)**	**(0.001)**	**(0.009)**
R Caudate	**−0.311**	**−0.378**	−0.249	**−0.347**	**−0.432**	**−0.347**	**−0.315**	**−0.413**	**−0.269**
	(0.008)	**(0.008)**	(0.066)	**(0.003)**	**(0.001)**	**(0.006)**	**(0.007)**	**(0.003)**	**(0.036)**
L NA	−0.142	−0.203	−0.117	**−0.272**	−0.266	−0.228	**−0.265**	**−0.336**	−0.129
	(0.239)	(0.141)	(0.399)	**(0.022)**	(0.067)	(0.122)	**(0.026)**	**(0.022)**	(0.282)
R NA	−0.085	−0.176	−0.055	**−0.369**	**−0.413**	**−0.359**	**−0.264**	**−0.366**	−0.197
	(0.481)	(0.204)	(0.681)	**(0.002)**	**(0.002)**	**(0.008)**	**(0.026)**	**(0.011)**	(0.122)
L Hippocampus	**−0.295**	**−0.391**	**−0.319**	**−0.352**	**−0.392**	**−0.259**	**−0.458**	**−0.467**	**−0.318**
	(0.012)	**(0.006)**	**(0.028)**	**(0.003)**	**(0.007)**	**(0.034)**	**(< 0.001)**	**(< 0.001)**	**(0.007)**
R Hippocampus	**−0.275**	**−0.394**	**−0.283**	**−0.487**	**−0.425**	**−0.337**	**−0.411**	**−0.419**	**−0.257**
	(0.020)	**(0.007)**	**(0.047)**	**(< 0.001)**	**(0.001)**	**(0.005)**	**(< 0.001)**	**(0.002)**	**(0.032)**
CC	**−0.335**	−0.231	−0.101	**−0.384**	**−0.292**	−0.228	**−0.400**	**−0.314**	−0.184
	(0.004)	(0.103)	(0.456)	**(0.001)**	**(0.036)**	(0.095)	**(0.001)**	**(0.023)**	(0.160)

NA nucleus accumbens; CC corpus callosum.
Bold print denotes significance at $p \leq 0.05$.
r Pearson correlation coefficient.
β_1 Standardised regression coefficient after controlling for sex of child, cigarettes/day during pregnancy, lead concentration (ug/dl), child age at scan and socioeconomic status.
β_2 Standardised regression coefficient after controlling for sex of child, cigarettes/day during pregnancy, lead concentration (ug/dl), child age at scan, socioeconomic status and TIV from [a]FreeSurfer v.6.0 or [b]FreeSurfer v.5.1.

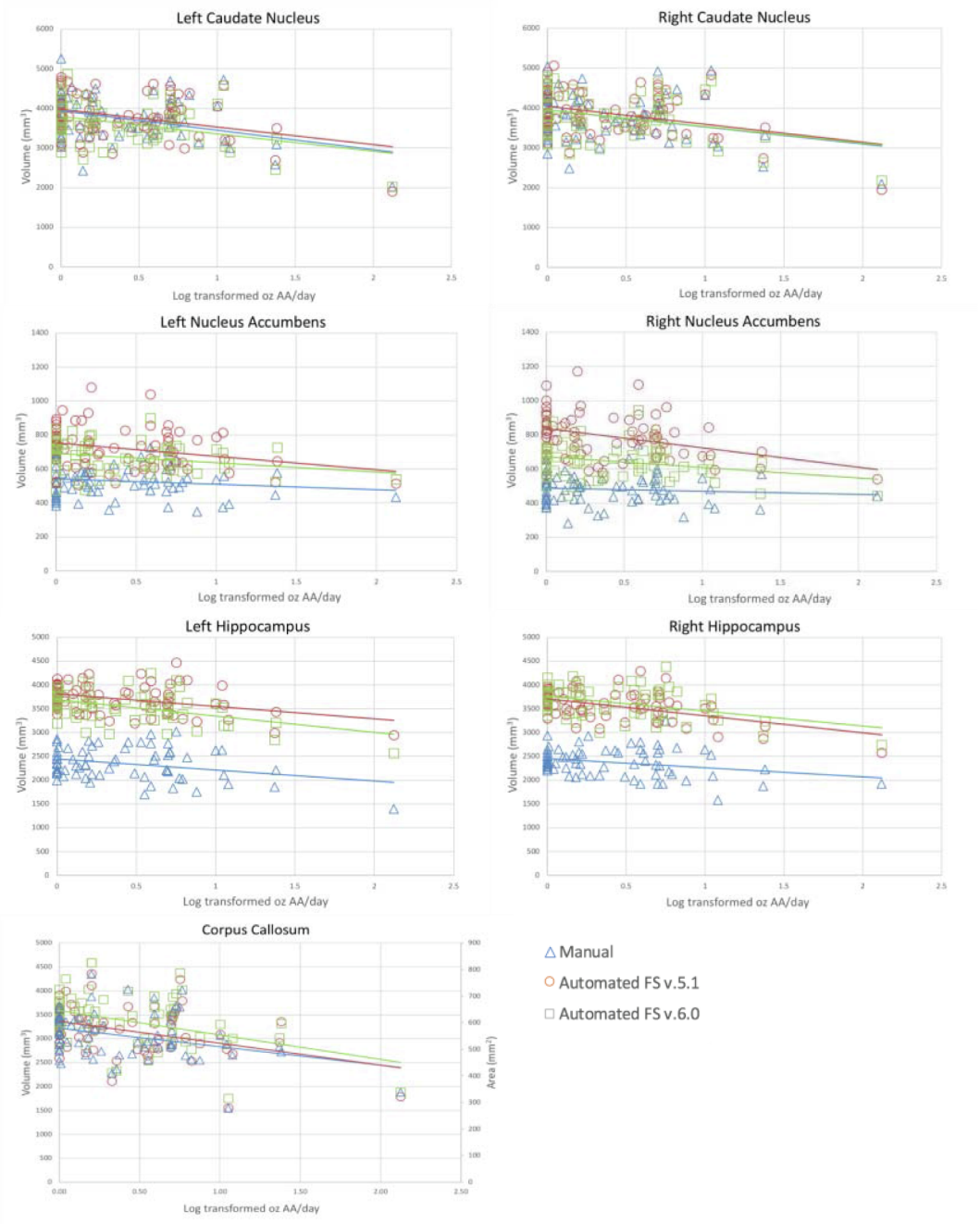

Fig. 3. Scatter plots for each region and each segmentation method of volume (and manually traced corpus callosum area) as a function of prenatal alcohol exposure.

this region. Of note (see Fig. 4B) is the erroneous inclusion of the anterior commissure (indicated by a white circle) by FreeSurfer v.5.1 in the volume of the left NA, which manual tracing and FreeSurfer v.6.0 correctly excluded. The small size of the NA may additionally exacerbate the already high variability and poor reliability in this region (Guenette et al., 2018; Lehmann et al., 2010; Schoemaker et al., 2016).

Conversely, the caudate nucleus has very clear borders, and delineation is well-defined and agreed upon in the literature (Archibald et al., 2001; Fryer et al., 2012; Looi et al., 2008). Intra- and inter-rater reliabilities of manual tracing were highest in this region, as were agreement and consistency across methods. Fig. 4 visually confirms that the delineation of the caudate nucleus by all three segmentation methods is very similar. Cronbach's α indicates that internal consistency across all three methods is greater than the desired standard of 0.95 in this region.

Comparisons of manual and automated segmentation have often

Table 5
Comparison for the three segmentation methods of the additional variance in the amount of prenatal alcohol exposure that is explained when hippocampal, caudal and CC sizes are added to the multiple regression model.

Segmentation Method	Model 1	Model 2		Model 3	
	R^2 change	R^2 change	p	R^2 change	p
Manual tracing[a]	0.342	0.050	**0.025**	0.071	0.188
FreeSurfer v5.1[b]	0.342	0.036	0.060	0.103	0.052
FreeSurfer v6.0[a]	0.342	0.050	**0.025**	0.112	**0.031**

Model 1 includes all confounders (child age, sex and lead; maternal smoking during pregnancy and socioeconomic status).
Model 2 includes confounders and TIV.
Model 3 includes confounders, TIV and ROI sizes (bilateral hippocampal and caudal volumes, and CC area/volume).
BOLD denotes significance at $p \leq 0.05$.
[a] TIV from FreeSurfer v6.0.
[b] TIV from FreeSurfer v5.1.

focused on the hippocampus, possibly due to the clinical implications of volume changes in this region (Akudjedu et al., 2018; Barnes et al., 2008; Dewey et al., 2010; Morey et al., 2009; Schoemaker et al., 2016). Although ICC consistency and internal consistency for the hippocampus were better than for the NA and reached the minimum acceptable level of 0.7, absolute agreement was the worst. Similar to findings from other studies, hippocampal volumes from both versions of FreeSurfer were larger than those from manual tracing (Akudjedu et al., 2018; Dewey et al., 2010; Schoemaker et al., 2016). Figs. 5 and 6 show that this overestimation is likely due to differences in segmentation protocols. Manual tracing did not include the alveous (indicated in Fig. 5A by the red arrow) or the fornices. Since this white matter is difficult to differentiate from the grey matter in the coronal plane, the hippocampus was manually traced in the sagittal plane (Fig. 6A and B) where the white matter is more easily visible in terms of intensity and anatomical location. In Fig. 5B–C and Fig. 6C–F it can be seen that both versions of FreeSurfer include the alveous, as well as random bits of the posterior horn of the lateral ventricle (black areas indicated by yellow arrows) in the hippocampus segmentation.

In Fig. 7, we highlight on a midline slice of one participant errors observed on the automated segmentations of the CC. Fig. 7B and C show that FreeSurfer erroneously includes some of the fornix (white arrow) in the posterior CC, and the artery curving around the CC anteriorly (red arrow) in the anterior and mid-anterior CC segmentations. In Fig. 7C, FreeSurfer v. 6.0 mislabelled the anterior and posterior CC (yellow arrows) as cerebral white matter.

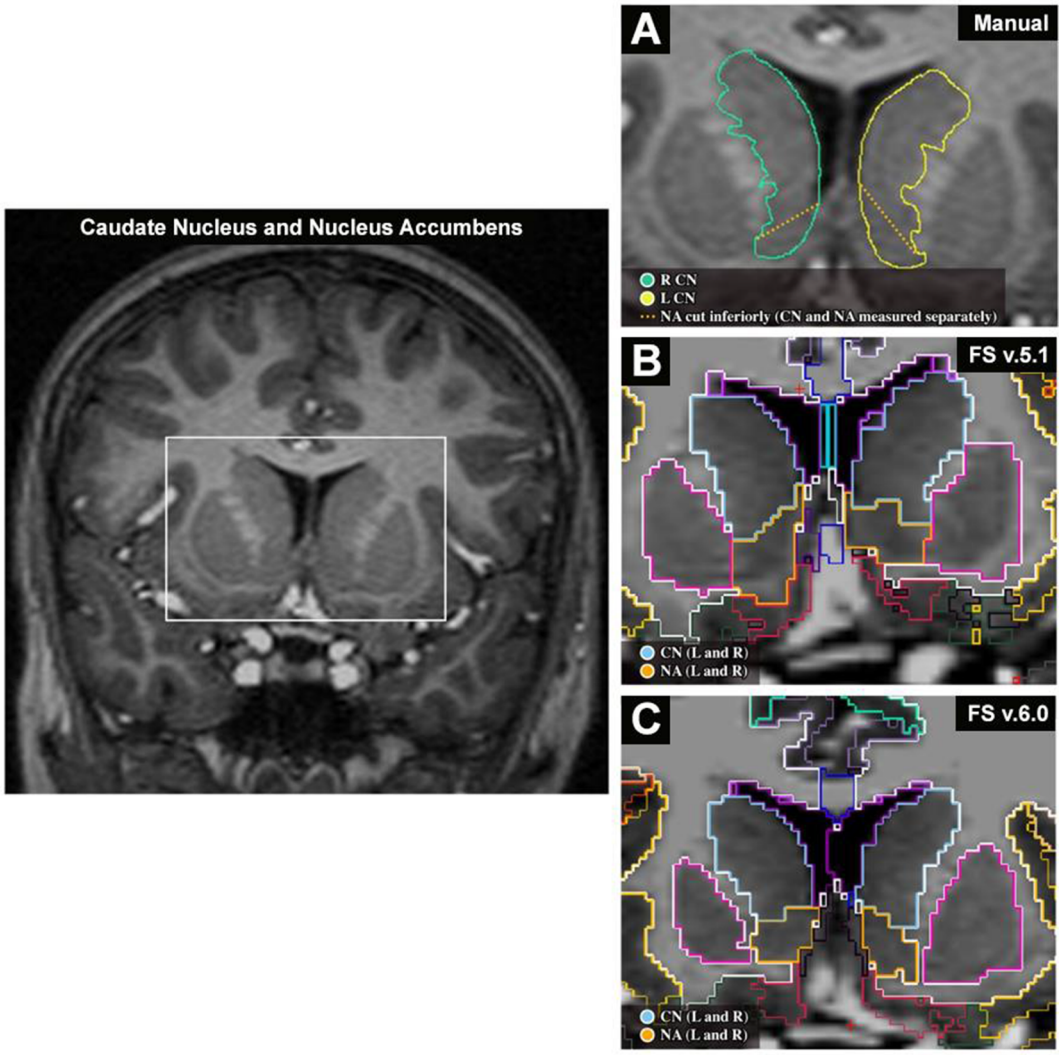

Fig. 4. (Left) Coronal slice showing the basal ganglia, and (right) zoomed images showing segmentation of the caudate nucleus (CN) and nucleus accumbens (NA) by (A) manual tracing, (B) FreeSurfer v. 5.1 and (C) FreeSurfer v. 6.0. The white circle in (B) highlights the erroneous inclusion of the anterior commissure by FreeSurfer v.5.1 in the left NA.

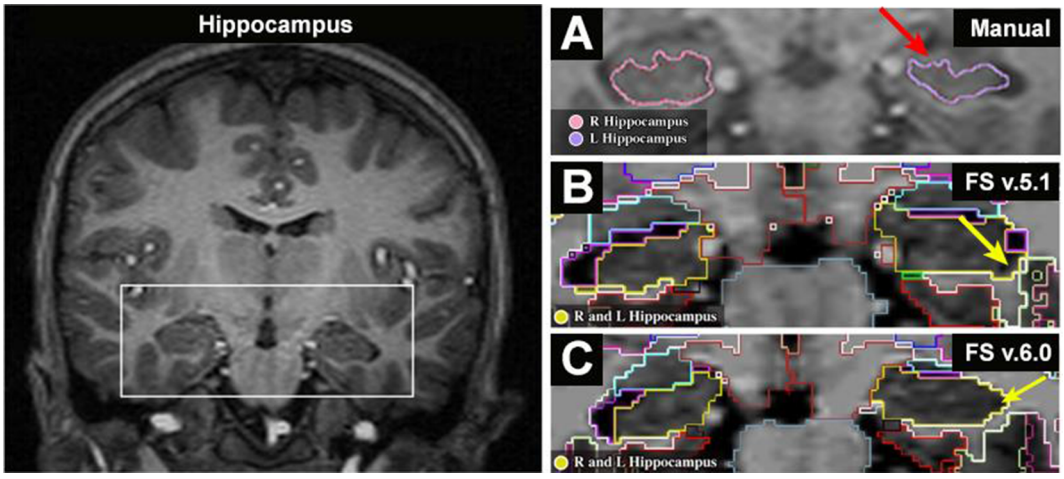

Fig. 5. (Left) Coronal slice showing the hippocampal region, and (right) zoomed images of the boxed region showing segmentation of the left and right hippocampus by (A) manual tracing, (B) FreeSurfer v. 5.1 and (C) FreeSurfer v. 6.0. The red arrow in (A) indicates the alveolus, and the yellow arrows in (B) and (C) show bits of the posterior horn of the lateral ventricle erroneously included in the FreeSurfer segmentations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Validity

As hypothesized, volumes from all three segmentation methods demonstrated that increasing PAE is associated with smaller CC, caudate nuclei and hippocampi. Notably, automated segmentation additionally demonstrated association of increasing PAE with smaller NA – findings that were not seen on manual tracing and that, for the most part, did not survive after control for TIV. Although automated segmentation demonstrated volumetric differences between diagnostic groups in more regions than manual tracing, both manual and FS6.0 showed *disproportionate* volume reductions, albeit below conventional levels of significance, in children with FAS/PFAS *only* in the right hippocampus. Volume reductions seen in other regions in children with FAS/PFAS on FreeSurfer v.6.0 did not survive after controlling for TIV, suggesting that these reductions may be a consequence of PAE-related reductions in overall brain size rather than a specific vulnerability of these regions to PAE.

Greater sensitivity in detecting dose-dependent than diagnostic PAE-related changes are consistent with findings in Biffen et al. (2018), and has been shown also for other imaging modalities, including functional MRI (Meintjes et al., 2014; du Plessis et al., 2015), brain metabolism (du Plessis et al., 2014) and cortical morphology (De Guio et al., 2014).

Due to the poor reliability across methods of NA segmentations, we were cautious in interpreting the associations of increasing PAE with smaller NA volumes seen *only* on automated segmentations as being 'real'. Notably, few of these associations (only right NA on FS5.1) remained significant after controlling for TIV. In contrast, associations seen in hippocampi, where reliability across all three methods on ICC consistency reached at least 0.70 (despite poor absolute agreement), were seen across all methods and survived after controlling both for potential confounders and TIV, thereby inspiring greater confidence in the validity of this result. It is a limitation of the present study that we are unable to say with absolute certainty which group differences or associations are "real" (whether or not they were consistently seen across methods), as we do not know the "actual" volumes of the brain regions being studied. Based on the findings from this study, we recommend that studies examining morphometric changes resulting from pathology should employ at least two different segmentation tools and that findings in regions demonstrating reliabilities across methods on ICC consistency less than 0.70 should be interpreted with caution.

Since manual tracing is typically regarded as the "gold standard" for volume measurement, it was unexpected that adding volumes from FS6.0 to the multiple regression model of PAE showed greater improvement than adding volumes from FS5.1 or manual tracing. This result, which may be due to reduced variability associated with automated segmentation compared to manual tracing that involves an element of subjectivity, points to greater sensitivity of FS6.0 in detecting PAE-related volume changes.

The findings of this study are likely not limited to a paediatric sample with PAE. The success of automated FreeSurfer segmentation (especially version 6.0) in detecting PAE-related effects in easy-to-segment regions and the hippocampus similar to those seen on manual tracing in this sample of 9- to 11-year-old children that differs greatly from the healthy Western adult-based sample that was used in the development of FreeSurfer (Fischl et al., 2002), suggests that it can be applied with relative confidence in these brain regions to geographically different clinical cohorts of this age. Previous work from our group on 5-year-old children, however, showed significant discrepancies between manual tracing and automated FreeSurfer segmentation in detecting HIV-related volumetric changes (Randall et al., 2017). Thus, care should be taken in younger populations where differences to an adult-based atlas and grey/white matter contrast differences due to incomplete myelination may be more significant.

5. Conclusion

The present work confirms excellent reliability of volume measurements using manual tracing and automated FreeSurfer segmentation in easy-to-segment regions such as the CC and caudate nucleus in 9- to 11-year-old children with and without PAE. Despite poor absolute agreement between methods in the NA and hippocampus, all three segmentation methods detected dose-dependent volume reductions in regions where reliabilities on ICC consistency across methods reached at least 0.70, namely the CC, and bilateral caudate nuclei and hippocampi. PAE-related changes in the NA for which ICC consistency did not reach this minimum were inconsistent across methods and should be interpreted with caution. This is the first study to demonstrate in a pre-adolescent cohort the ability of automated segmentation with FreeSurfer to detect regional volume changes associated with pathology in the same regions, in the same direction and of the same order of magnitude as found using manual tracing.

Author contributions

As part of her doctoral dissertation, SB performed the tracing and

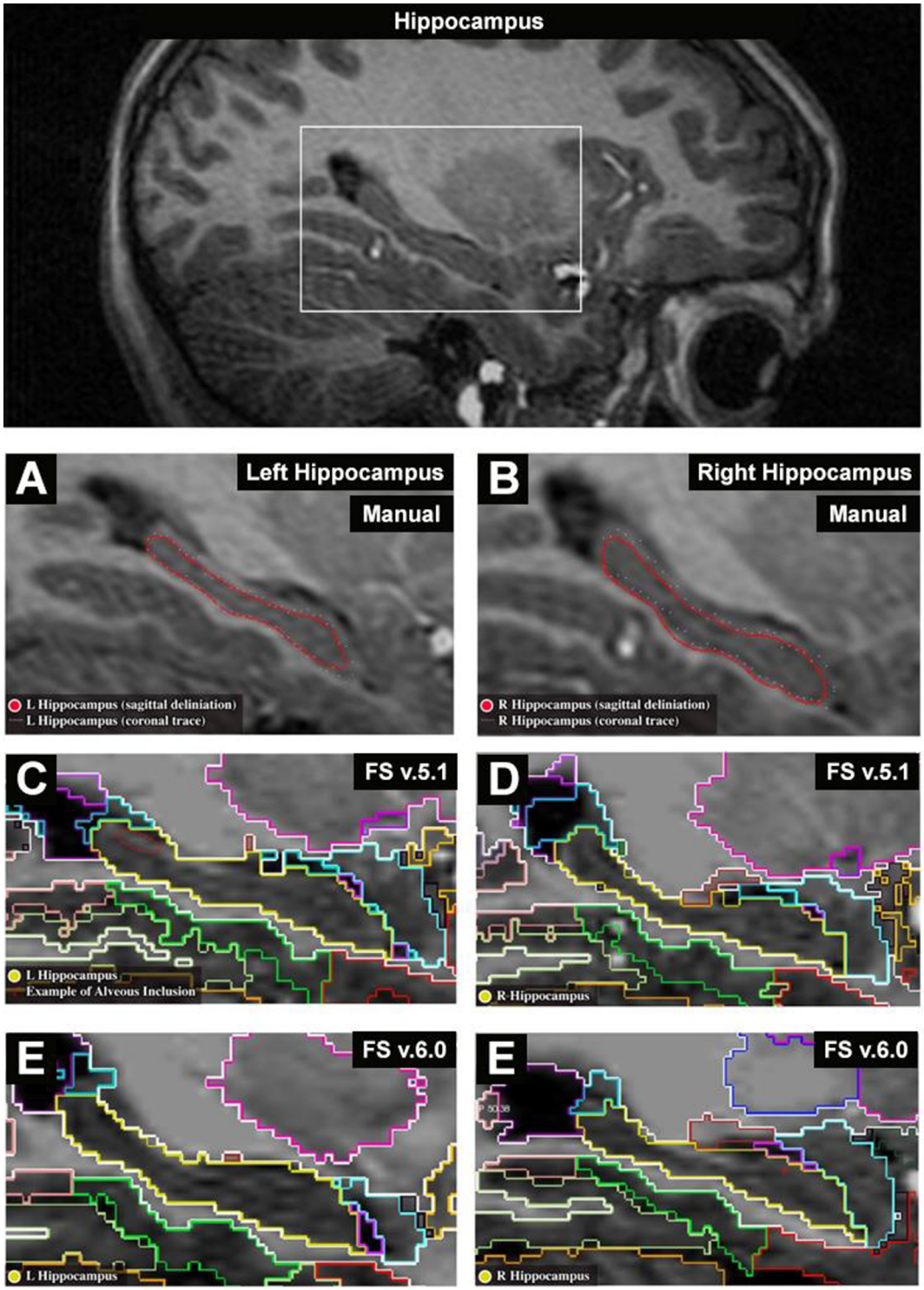

Fig. 6. (Top) Hippocampal region on a sagittal slice, and zoomed images of the boxed region showing segmentation by (A,B) manual tracing, (C,D) FreeSurfer v.5.1, and (E,F) FreeSurfer v.6.0 for the (left column) left and (right column) right hippocampus, respectively.

segmentation of the volumes under the supervision of CW. She reviewed the literature, performed the data analyses, and interpreted and wrote up the findings. EM provided overall project supervision, collaborated on the design of the neuroimaging study, the data analysis, interpretation and write-up of the findings. SJ and JJ designed the original FASD study, supervised recruitment of the cohort and maternal interviews and child assessments, collaborated on the data analysis, the interpretation of the findings and write-up of the paper. CM administered the maternal interviews, which included sociodemographic information and alcohol, smoking and drug ascertainment. ND performed additional data analyses and data representation.

Funding

This study was supported by NIH/NIAAA grants R01AA016781 and U01AA014790; National Research Foundation (NRF) of South Africa (Grant number 48337); Medical Research Council of South Africa; and the Lycaki-Young Fund, State of Michigan.

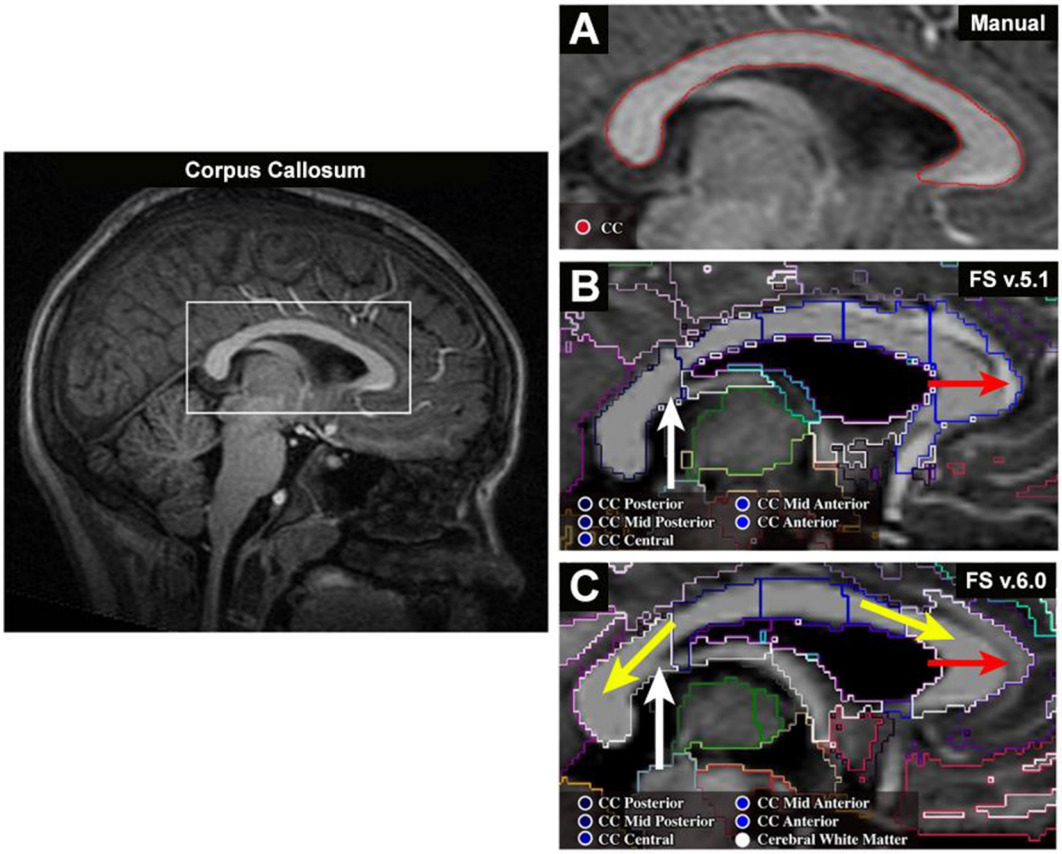

Fig. 7. (Left) Mid-sagittal slice showing the corpus callosum (CC), and (right) zoomed images of the boxed region showing segmentation of the CC by (A) manual tracing, (B) FreeSurfer v. 5.1 and (C) FreeSurfer v. 6.0. The white arrows in (B) and (C) indicate the fornix and the red arrows the artery curving around the CC anteriorly that are often erroneously included in the CC segmentation. The yellow arrows in (C) indicate anterior and posterior sections of the CC that were mislabelled as cerebral white matter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Conflict of interest

The authors declare no competing financial interests.

Acknowledgments

We thank the dysmorphologists H. Eugene Hoyme, M.D., Luther K. Robinson, M.D., and Nathaniel Khaole, M.D., who examined the children in our 2005 FASD diagnostic clinic, Drs. Hoyme and Robinson in the 2009 clinic, and Dr. Hoyme who led the team of dysmorphologists in the 2013 and 2016 clinics. We thank R. Colin Carter, M.D./M.M.Sc. for his consultation and participation in the clinics; the CUBIC radiographers Marie-Louise de Villiers and Nailah Maroof; Steven Randall (S.R.) for his hand tracing work for the inter-rater reliability assessments; and our University of Cape Town and Wayne State University research staff Nicolette Hamman, Mariska Pienaar, Maggie September, Emma Makin, Nadine Lindinger, Catherine Lewis, Beverly Arendse, and Renee Sun. We also thank the parents and children for their long-term participation in and contribution to the study.

References

Akhondi-Asl, A., Jafari-Khouzani, K., Elisevich, K., Soltanian-Zadeh, H., 2011. Hippocampal volumetry for lateralization of temporal lobe epilepsy: automated versus manual methods. Neuroimage 54, S218–S226. https://doi.org/10.1016/J.NEUROIMAGE.2010.03.066.

Akudjedu, T.N., Nabulsi, L., Makelyte, M., Scanlon, C., Hehir, S., Casey, H., Ambati, S., Kenney, J., O'Donoghue, S., McDermott, E., Kilmartin, L., Dockery, P., McDonald, C., Hallahan, B., Cannon, D.M., 2018. A comparative study of segmentation techniques for the quantification of brain subcortical volume. Brain Imaging Behav. https://doi.org/10.1007/s11682-018-9835-y.

Archibald, S.L., Fennema-Notestine, C., Gamst, A., Riley, E.P., Mattson, S.N., Jernigan, T.L., 2001. Brain dysmorphology in individuals with severe prenatal alcohol exposure. Dev. Med. Child Neurol. 43, 148–154. https://doi.org/10.1111/j.1469-8749.2001.tb00179.x.

Barnes, J., Foster, J., Boyes, R.G., Pepple, T., Moore, E.K., Schott, J.M., Frost, C., Scahill, R.I., Fox, N.C., 2008. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. https://doi.org/10.1016/j.neuroimage.2008.01.012.

Biffen, S.C., Warton, C.M.R., Lindinger, N.M., Randall, S.R., Lewis, C.E., Molteno, C.D., Jacobson, J.L., Jacobson, S.W., Meintjes, E.M., 2018. Reductions in corpus callosum volume partially mediate effects of prenatal alcohol exposure on IQ. Front. Neuroanat. 11, 132. https://doi.org/10.3389/fnana.2017.00132.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J.G., de Leon, M.J., deToledo-Morrell, L., Killiany, R.J., Lehéricy, S., Pantel, J., Pruessner, J.C., Soininen, H., Watson, C., Duchesne, S., Jack Jr, C.R., Frisoni, G.B., 2011. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J. Alzheimer's Dis. 26, 61–75. https://doi.org/10.3233/JAD-2011-0004.

Cardenas, V.A., Price, M., Infante, M.A., Moore, E.M., Mattson, S.N., Riley, E.P., Fein, G., 2014. Automated cerebellar segmentation: validation and application to detect smaller volumes in children prenatally exposed to alcohol. NeuroImage Clin. 4, 295–301. https://doi.org/10.1016/j.nicl.2014.01.002.

Carmines, E., Zeller, R., 1979. Reliability and validity assessment. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America. https://doi.org/10.4135/9781412985642.

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, e5265. https://doi.org/10.1371/journal.pone.0005265.

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd edition. Lawrence Earlbaum Associates, Mahwah, NJ.

De Guio, François, Mangin, Jean-François, Rivière, Denis, Perrot, Matthieu, Molteno, Christopher, Jacobson, Sandra, Meintjes, Ernesta, Jacobson, Joseph, 2014. A study of cortical morphology in children with fetal alcohol spectrum disorders. Human Brain Map. 35 (5), 2285–2296. https://doi.org/10.1002/hbm.22327.

Dewey, J., Hana, G., Russell, T., Price, J., McCaffrey, D., Harezlak, J., Sem, E., Anyanwu, J.C., Guttmann, C.R., Navia, B., Cohen, R., Tate, D.F., 2010. Reliability and validity of

MRI-based automated volumetry software relative to auto-assisted manual measurement of subcortical structures in HIV-infected patients from a multisite study. Neuroimage 51, 1334–1344. https://doi.org/10.1016/j.neuroimage.2010.03.033.
Doring, T.M., Kubo, T.T.A., Domingues, R.C., Gasparetto, E.L., 2010. Evaluation of hippocampal volume based on MRI applying manual and automatic segmentation techniques. Rev. Bras. Física Médica 4, 89–91. https://doi.org/10.29384/RBFM.2010.V4.N1.P89-91.
du Plessis, Lindie, Jacobson, Joseph, Jacobson, Sandra, Hess, Aaron, van der Kouwe, Andre, Avison, Malcolm, Molteno, Christopher, Stanton, Mark, Stanley, Jeffrey, Peterson, Bradley, Meintjes, Ernesta, 2014. An in vivo 1H magnetic resonance spectroscopy study of the deep cerebellar nuclei in children with fetal alcohol spectrum disorders. Alcohol.: Clin. Exp. Res. 38 (5), 1330–1338. https://doi.org/10.1111/acer.12380.
du Plessis, Lindie, Jacobson, Sandra, Molteno, Christopher, Robertson, Frances, Peterson, Bradley, Jacobson, Joseph, Meintjes, Ernesta, 2015. Neural correlates of cerebellar-mediated timing during finger tapping in children with fetal alcohol spectrum disorders. NeuroImage: Clinical 7, 562–570. https://doi.org/10.1016/j.nicl.2014.12.016.
Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X.
Fryer, S.L., Mattson, S.N., Jernigan, T.L., Archibald, S.L., Jones, K.L., Riley, E.P., 2012. Caudate volume predicts neurocognitive performance in youth with heavy prenatal alcohol exposure. Alcohol. Clin. Exp. Res. 36, 1932–1941. https://doi.org/10.1111/j.1530-0277.2012.01811.x.
Gronenschild, E.H.B.M., Habets, P., Jacobs, H.I.L., Mengelers, R., Rozendaal, N., van Os, J., Marcelis, M., 2012. The effects of freeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS ONE 7, e38234. https://doi.org/10.1371/journal.pone.0038234.
Guenette, J.P., Shenton, M.E., Koerte, I.K., 2018. Imaging of concussion in young athletes. Neuroimaging Clin. N. Am. 28, 43–53. https://doi.org/10.1016/J.NIC.2017.09.004.
Helms, G., 2016. Segmentation of human brain using structural MRI. Magn. Reson. Mater. Physics Biol. Med. 29, 111–124. https://doi.org/10.1007/s10334-015-0518-z.
Hollingshead, A.B., 2011. Four factor index of social status. Yale J. Sociol. 8, 21–51.
Hoyme, H.E., May, P.A., Kalberg, W.O., Kodituwakku, P., Gossage, J.P., Trujillo, P.M., Buckley, D.G., Miller, J.H., Aragon, A.S., Khaole, N., Viljoen, D.L., Jones, K.L., Robinson, L.K., 2005. A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: clarification of the 1996 institute of medicine criteria. Pediatrics 115, 39–47. https://doi.org/10.1542/peds.2004-0259.
Jacobson, S.W., Chiodo, L.M., Sokol, R.J., Jacobson, J.L., 2002. Validity of maternal report of prenatal alcohol, cocaine, and smoking in relation to neurobehavioral outcome. Pediatrics 109, 815–825. https://doi.org/10.1542/peds.109.5.815.
Jacobson, S.W., Stanton, M.E., Molteno, C.D., Burden, M.J., Fuller, D.S., Hoyme, H.E., Robinson, L.K., Khaole, N., Jacobson, J.L., 2008. Impaired eyeblink conditioning in children with fetal alcohol syndrome. Alcohol. Clin. Exp. Res. 32, 365–372. https://doi.org/10.1111/j.1530-0277.2007.00585.x.
Lance, C.E., Butts, M.M., Michels, L.C., 2006. The sources of four commonly reported cutoff criteria. Organ. Res. Methods 9, 202–220. https://doi.org/10.1177/1094428105284919.
Lehmann, M., Douiri, A., Kim, L.G., Modat, M., Chan, D., Ourselin, S., Barnes, J., Fox, N.C., 2010. Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements. Neuroimage 49, 2264–2274. https://doi.org/10.1016/J.NEUROIMAGE.2009.10.056.
Looi, J.C.L., Lindberg, O., Liberg, B., Tatham, V., Kumar, R., Maller, J., Millard, E., Sachdev, P., Högberg, G., Pagani, M., Botes, L., Engman, E.-L., Zhang, Y., Svensson, L., Wahlund, L.-O., 2008. Volumetrics of the caudate nucleus: reliability and validity of a new manual tracing protocol. Psychiatry Res. 163, 279–288. https://doi.org/10.1016/j.pscychresns.2007.07.005.
May, P.A., Blankenship, J., Marais, A.-S., Gossage, J.P., Kalberg, W.O., Barnard, R., De Vries, M., Robinson, L.K., Adnams, C.M., Buckley, D., Manning, M., Jones, K.L., Parry, C., Hoyme, H.E., Seedat, S., 2013. Approaching the prevalence of the full spectrum of fetal alcohol spectrum disorders in a South African population-based study. Alcohol. Clin. Exp. Res. 37, 818–830. https://doi.org/10.1111/acer.12033.
Meintjes, E.M., Narr, K.L., der Kouwe, A.J.W. van, Molteno, C.D., Pirnia, T., Gutman, B., Woods, R.P., Thompson, P.M., Jacobson, J.L., Jacobson, S.W., 2014. A tensor-based morphometry analysis of regional differences in brain volume in relation to prenatal alcohol exposure. NeuroImage Clin. 5, 152–160. https://doi.org/10.1016/j.nicl.2014.04.001.
Morey, R.A., Petty, C.M., Xu, Y., Hayes, J.P., Wagner, H.R., Lewis, D.V., LaBar, K.S., Styner, M., McCarthy, G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45, 855–866. https://doi.org/10.1016/j.neuroimage.2008.12.033.
Mulder, E.R., de Jong, R.A., Knol, D.L., van Schijndel, R.A., Cover, K.S., Visser, P.J., Barkhof, F., Vrenken, H., 2014. Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 92, 169–181. https://doi.org/10.1016/J.NEUROIMAGE.2014.01.058.
Nugent, A.C., Luckenbaugh, D.A., Wood, S.E., Bogers, W., Zarate, C.A., Drevets, W.C., 2013. Automated subcortical segmentation using FIRST: test-retest reliability, interscanner reliability, and comparison to manual segmentation. Hum. Brain Mapp. 34, 2313–2329. https://doi.org/10.1002/hbm.22068.
Nunnally, J.C., Bernstein, I.H., Berge, J.M.T., 1967. Psychometric Theory, 3rd ed. McGraw-Hill, New York.
Pardoe, H.R., Pell, G.S., Abbott, D.F., Jackson, G.D., 2009. Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? Epilepsia 50, 2586–2592. https://doi.org/10.1111/j.1528-1167.2009.02243.x.
Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56, 907–922. https://doi.org/10.1016/J.NEUROIMAGE.2011.02.046.
Pipitone, J., Park, M.T.M., Winterburn, J., Lett, T.A., Lerch, J.P., Pruessner, J.C., Lepage, M., Voineskos, A.N., Chakravarty, M.M., 2014. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. Neuroimage 101, 494–512. https://doi.org/10.1016/J.NEUROIMAGE.2014.04.054.
Randall, S.R., Warton, C.M.R., Holmes, M.J., Cotton, M.F., Laughton, B., van der Kouwe, A.J.W., Meintjes, E.M., 2017. Larger subcortical gray matter structures and smaller corpora callosa at age 5 years in HIV infected children on early ART. Front. Neuroanat. 11, 95. https://doi.org/10.3389/fnana.2017.00095.
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 Res. Neuroimaging 181, 219–225. https://doi.org/10.1016/j.pscychresns.2009.10.011.
Schoemaker, D., Buss, C., Head, K., Sandman, C.A., Davis, E.P., Chakravarty, M.M., Gauthier, S., Pruessner, J.C., 2016. Hippocampus and amygdala volumes from magnetic resonance images in children: assessing accuracy of FreeSurfer and FSL against manual segmentation. Neuroimage 129, 1–14. https://doi.org/10.1016/j.neuroimage.2016.01.038.
Shen, L., Kim, S., Risacher, S.L., Nho, K., Swaminathan, S., West, J.D., Foroud, T., Pankratz, N., Moore, J.H., Sloan, C.D., Huentelman, M.J., Craig, D.W., DeChairo, B.M., Potkin, S.G., Jack, C.R., Weiner, M.W., Saykin, A.J., 2010. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. Neuroimage 53, 1051–1063. https://doi.org/10.1016/J.NEUROIMAGE.2010.01.042.
Soper, D.S., 2020. Significance of the Difference between Two Slopes Calculator [Software]. Available from http://www.danielsoper.com/statcalc.
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. Neuroradiology 50, 569–581. https://doi.org/10.1007/s00234-008-0383-9.
Terwee, C.B., Bot, S.D.M., de Boer, M.R., van der Windt, D.A.W.M., Knol, D.L., Dekker, J., Bouter, L.M., de Vet, H.C.W., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60, 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.
Tisdall, M.D., Hess, A.T., Reuter, M., Meintjes, E.M., Fischl, B., van der Kouwe, A.J.W., 2012. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. Magn. Reson. Med. 68, 389–399. https://doi.org/10.1002/mrm.23228.
van der Kouwe, A.J.W., Benner, T., Salat, D.H., Fischl, B., 2008. Brain morphometry with multiecho MPRAGE. Neuroimage 40, 559–569. https://doi.org/10.1016/j.neuroimage.2007.12.025.
Woods, R.P., 2003. Multitracer: a Java-based tool for anatomic delineation of grayscale volumetric images. Neuroimage 19, 1829–1834. https://doi.org/10.1016/S1053-8119(03)00243-X.
Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M.C., Kaus, M.R., Haker, S.J., Wells, W.M., Jolesz, F.A., Kikinis, R., Kikinis, R., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Acad. Radiol. 11, 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8.