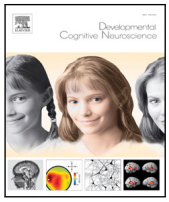
Developmental Cognitive Neuroscience 52 (2021) 101028

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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



Next-gen tools

Charting the human amygdala development across childhood and adolescence: Manual and automatic segmentation

Quan Zhou [a,b], Siman Liu [a,b], Chao Jiang [c], Ye He [d], Xi-Nian Zuo [e,a,b,f,g,*], The Chinese Color Nest Consortium

[a] Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China
[b] Department of Psychology, University of Chinese Academy of Sciences, Beijing, 100049, China
[c] School of Psychology, Capital Normal University, Beijing, 100048, China
[d] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[e] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, 100875, China
[f] National Basic Science Data Center, Beijing, 100190, China
[g] Developmental Population Neuroscience Research Center, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875, China

ARTICLE INFO

Keywords:
Amygdala
Brain development
Growth chart
MRI
Reliability

ABSTRACT

The developmental pattern of the amygdala throughout childhood and adolescence has been inconsistently reported in previous neuroimaging studies. Given the relatively small size of the amygdala on full brain MRI scans, discrepancies may be partly due to methodological differences in amygdalar segmentation. To investigate the impact of volume extraction methods on amygdala volume, we compared *FreeSurfer*, *FSL* and *volBrain* segmentation measurements with those obtained by manual tracing. The manual tracing method, which we used as the 'gold standard', exhibited almost perfect intra- and inter-rater reliability. We observed systematic differences in amygdala volumes between automatic (*FreeSurfer* and *volBrain*) and manual methods. Specifically, compared with the manual tracing, *FreeSurfer* estimated larger amygdalae, and *volBrain* produced smaller amygdalae while *FSL* demonstrated a mixed pattern. The tracing bias was not uniform, but higher for smaller amygdalae. We further modeled amygdalar growth curves using accelerated longitudinal cohort data from the Chinese Color Nest Project (http://deepneuro.bnu.edu.cn/?p=163). Trajectory modeling and statistical assessments of the manually traced amygdalae revealed linearly increasing and parallel developmental patterns for both girls and boys, although the amygdalae of boys were larger than those of girls. Compared to these trajectories, the shapes of developmental curves were similar when using the *volBrain* derived volumes. *FreeSurfer* derived trajectories had more nonlinearities and appeared flatter. *FSL* derived trajectories demonstrated an inverted U shape and were significantly different from those derived from manual tracing method. The use of amygdala volumes adjusted for total gray-matter volumes, but not intracranial volumes, resolved the shape discrepancies and led to reproducible growth curves between manual tracing and the automatic methods (except *FSL*). Our findings revealed steady growth of the human amygdala, mirroring its functional development across the school age. Methodological improvements are warranted for current automatic tools to achieve more accurate amygdala structure at school age, calling for next generation tools.

1. Introduction

Childhood and adolescence are key periods for socioemotional development, which correlate strongly with the development of risk factors for diverse neuropsychiatric disorders (Paus et al., 2008). Together with enhanced efforts to prevent such disorders, many large-scale studies have been undertaken to explore behavioral and biological development of children and adolescents (Ortiz and Raine, 2004; Silk et al., 2007; Connor, 2004). Rapid progress in in-vivo brain imaging technologies has accelerated the use of structural magnetic resonance imaging (MRI) to quantify volumes of different brain structures. These morphological features have been demonstrated by MRI to be sensitive for developmental brain changes (Tamnes et al., 2013; Albaugh et al., 2017; Wierenga et al., 2018). The accurate establishment of developmental trajectories of brain structures using MRI is thus an important

* Corresponding author at: Developmental Population Neuroscience Research Center, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875, China.
E-mail addresses: xinian.zuo@bnu.edu.cn, zuoxn@psych.ac.cn (X.-N. Zuo).

https://doi.org/10.1016/j.dcn.2021.101028
Received 1 April 2021; Received in revised form 20 August 2021; Accepted 19 October 2021
Available online 28 October 2021

requirement for understanding the neurodevelopmental mechanisms of these disorders occurring during childhood and adolescence.

The amygdala is an almond-shaped brain structure, part of the limbic system and is highly connected with other brain regions (Schumann and Amaral, 2005). It plays important roles in emotional and cognitive processes, especially fear and threat processing (LeDoux, 1998; Cardinal et al., 2002; Pessoa, 2010) and exhibits network-level connectivity changes across the human lifespan (He et al., 2016). Abnormal amygdalar structure in children and adolescents has been related to a plethora of neurodevelopmental abnormalities (Scherf et al., 2013; Schumann et al., 2011), including autism (Mosconi et al., 2009; Schumann et al., 2004), anxiety disorder (De Bellis et al., 2000; Redlich et al., 2015) and schizophrenia (Ganzola et al., 2014). Meanwhile, many studies have explored age-related changes of the amygdala in pediatric samples (Uematsu et al., 2012; Gilmore et al., 2012; Wierenga et al., 2014; Barnea-Goraly et al., 2014; Herting et al., 2018), indicating the promise of using normal growth patterns for monitoring abnormal development. Growth charts are expected to aid risk evaluation, early diagnosis and educational monitoring by delineating typical development standards. In several recent studies, researchers have tracked the age-related increases of amygdala volume from childhood through adolescence (Herting et al., 2018; Goddings et al., 2014; Albaugh et al., 2017). However, a study including 271 individuals aged 8–29 years reported no significant changes in amygdala volume (Wierenga et al., 2018). This was similar to the observation from a sample of 85 individuals scanned twice across 8–22 years (Tamnes et al., 2013). Thus, there are mixed findings in the literature related to age-related differences or changes in amygdala volume. The anatomical complexity can limit the accurate measurement of amygdalar volume, leading to a large variation in findings obtained using different amygdala segmentation methods (Lyden et al., 2016), which may explain this inconsistency and reproducibility issue (Mills and Tamnes, 2014; Lyden et al., 2016).

Manual tracing is commonly considered as the 'gold standard' for amygdala segmentation (Morey et al., 2009). It enables flexible quantification guided by prior anatomical knowledge, without the need to make any of the assumptions built into algorithms. Experienced human tracers can correctly label ambiguous borders by adjusting for variation caused by complex or atypical anatomy and image artifacts. To increase reliability and reduce potential biases associated with manual tracing, multiple protocols have been generated and described in the literature (Schumann et al., 2004; Pruessner et al., 2000; Watson et al., 1992). These protocols significantly increase intra- and inter-rater agreement (Pruessner et al., 2000). However, manual tracing is time-consuming and requires the operator to have sufficient anatomical expertise. For large MRI datasets, the labor cost of manual tracing is prohibitive (Akudjedu et al., 2018; Schmidt et al., 2018). There is also subtly drift in tracing criteria of manual raters during the course of a long study. Accordingly, it is critical to develop automatic techniques that can accurately segment amygdala structures from large and growing datasets while providing consistent results and minimizing the human effort necessary for manual tracing.

Several tools have been developed to achieve automatic segmentation in a time-efficient manner including *FSL*-FIRST, *FreeSurfer* and *volBrain*, which are both freely available, ease to use, nearly fully automatic, and very accurate (Fischl et al., 2002; Manjón and Coupé, 2016; Morey et al., 2009; Akudjedu et al., 2018; Schmidt et al., 2018; Næss-Schmidt et al., 2016). FIRST was provided as part of the *FSL* software library to estimate boundaries of brain structures based on the signal intensity of the T1 image as well as the expected shape of structures using a probabilistic framework (Patenaude et al., 2011). *FreeSurfer* automatically assigns a label to each voxel from anatomical images based on probabilistic estimations relying on Markov random fields (Fischl et al., 2002). It may be difficult for this model-based method applied in *FreeSurfer* to model the regions of interest with sufficient accuracy in highly variable MRI data, such as inter-individual differences or pathological changes in neuroanatomy. To address this, multi-atlas label fusion approach such as *volBrain* has also been implemented. Multi-atlas label fusion segmentation techniques could combine multiple atlas information, thereby minimizing mislabeling from inaccurate affine or non-linear registration (Manjón and Coupé, 2016). Although automated segmentation has been shown to be comparable to manual tracing for adult populations (Fischl et al., 2002; Manjón and Coupé, 2016; Morey et al., 2009; Grimm et al., 2015), its performance for child and adolescent samples, in which head size and shape as well as the pace of structural growth differ, has not been validated adequately (Herten et al., 2019). In addition, the effects of any differences in the accuracy of automatic and manual amygdala segmentation on the subsequent examination of amygdala development in school-age children and adolescents remain incompletely understood.

To fully characterize similarities and discrepancies among techniques, we compared amygdala volumes obtained manually to those extracted by *FreeSurfer*, *volBrain* and FIRST in *FSL* using 427 longitudinal structural MRI scans from 198 healthy children and adolescents (baseline age: 6–17 years). To answer the aforementioned question, we examined how different tracing methods lead to trajectory differences in amygdala development across school age. Based upon previous reports (Morey et al., 2009; Schoemaker et al., 2016), we expected to observe systematic differences in amygdala segmentation performance among the three tracing methods. We hypothesized that such differences would affect the modeling of human amygdala growth.

Table 1
Sample characteristics for each time-point.

	Wave 1	Wave 2	Wave 3
n	183	149	95
n females/males	100/83	75/74	48/47
Age,	11.82	12.33	12.77
mean (SD)	(3.14)	(2.87)	(2.61)
Age, range	6–17	7–18	9–19

Table 2
Intra- and inter-rater reliability for rater QZ and rater ZQZ.

Reliability type	Rater	Hemisphere	ICC	95% Confidence interval	
				Lower bound	Upper bound
Intra-rater reliability	Rater QZ	Left	0.95	0.89	0.97
		Right	0.94	0.86	0.97
	Rater ZQZ	Left	0.91	0.82	0.96
		Right	0.91	0.82	0.96
Inter-rater reliability	Between rater QZ and ZQZ	Left	0.88	0.80	0.96
		Right	0.89	0.83	0.95

2. Materials and methods

2.1. Participants

The sample described in this study was part of an accelerated longitudinal database, namely the Chinese Color Nest Project (CCNP: http://deepneuro.bnu.edu.cn/?p=163) for developmental brain–mind association studies across different stages of the postnatal lifespan (Zuo et al., 2017; Liu et al., 2021). Such acceleration was implemented by combining cross-sectional and longitudinal design to achieve long-time follow-up studies, such as lifespan development cohorts (Nooner et al., 2012; Thompson et al., 2011). It was part of the developmental component of CCNP (devCCNP), and collected at Southwest University (devCCNP-SWU), Chongqing, China. The devCCNP-SWU was designed to delineate normative trajectories of brain development in the Chinese population across the school-aged years. The participants had no neurological or mental health problem and did not use psychotropic medication; their estimated intelligence quotients were $\geqslant$ 80. The devCCNP-SWU samples included data from 201 typically developing

controls (TDCs) aged 6–17 years who were invited to participate in three consecutive waves of data collection at intervals of approximately 1.25 years (Dong et al., 2020, 2021). T1-weighted MRI examinations were performed at these time points, and the images were visually inspected to exclude those with substantial head-motion artifacts and those with structural abnormalities. After this initial quality control, the final sample included 427 scans from 198 participants (105 females; 93 males; Table 1). Scans from three time points, two time points, and one time point were available for 79, 71, and 48 participants, respectively. The mean number of scans per participant was 2.16 (standard deviation = 0.79). The current study was approved by review committees of the participating institutions (Institute of Psychology, Chinese Academy of Sciences, and Southwest University).

2.2. MRI acquisition

All participants underwent MRI examinations performed with a Siemens Trio™ 3.0 T MRI scanner. A high-resolution magnetization-prepared rapid gradient-echo (MP-RAGE) T1 sequence (matrix = 256 × 256, FOV = 256 × 256 mm^2, slices thickness = 1 mm, repetition time (TR) = 2600 ms, echo time (TE) = 3.02 ms, inversion time (TI) = 900 ms, flip angle = 15°, number of slices = 176) was obtained for each individual.

2.3. Volumetric MRI preprocessing and segmentation

All the images were anonymized by removing all the personal information from the raw MRI data. We removed the facial information by using the **facemasking** tool (Milchenko and Marcus, 2013) customized with the Chinese pediatric templates developed by our lab (Dong et al., 2020), which has been integrated into the Connectome Computation System (Xu et al., 2015). The anonymized images were then uploaded to the online image processing system *volBrain* (http://volbrain.upv.es) for brain extraction (Manjón and Coupé, 2016). All the extracted individual brains were also denoised by spatially adaptive non-local means and corrected for intensity normalization in *volBrain*. These preprocessed brain volumes were all in the native space and ready for subsequent manual and automatic tracing procedures. Of note, we confirmed that the impacts of the face masking on the brain extraction and preprocessing are trifling by checking the individual images.

2.3.1. Manual tracing and reliability assessment

Anatomically trained raters QZ (the first author Quan Zhou) and ZQZ performed manual amygdala segmentation in the native space using the *ITK-SNAP* software (ver. 3.8.0) (Yushkevich et al., 2006). The anatomical boundaries of amygdala structures were defined and segmented according to the protocol described by Pruessner et al. (2000). This protocol has been demonstrated to achieve almost perfect intra- and inter-rater reliability. The reliability was quantified with intraclass correlation coefficient (ICC), which was interpreted as indicating slight $[0, 0.20)$, fair $[0.20, 0.40)$, moderate $[0.40, 0.60)$, substantial $[0.60, 0.80)$, or almost perfect $[0.80, 1]$ reliability (Landis and Koch, 1977; Xing and Zuo, 2018). To assess reliability for the protocol implementation in this study, QZ and ZQZ independently traced the amygdala volumes of 30 scans twice at a two-week interval. They were chosen from 30 subjects at baseline examination balanced for age and sex. The ICCs with a 95% confidence interval (CI) are derived by the following hierarchical linear mixed model on the repeated tracing volumes

$$\begin{aligned} V_{ijk} = \quad & \gamma_{000} + \text{subject}_{i00} + \text{order}_{j} + \text{rater}_{k} \\ & + \text{subject} \times \text{order}_{ij} \\ & + \text{subject} \times \text{rater}_{ik} \\ & + \text{order} \times \text{rater}_{jk} \\ & + e_{ijk} \end{aligned} \tag{1}$$

where V_{ijk} represents the amygdalar volume measurement for the ith ($i = 1, 2, \ldots, 30$) participant in the jth ($j = 1, 2$) manual tracing by the kth rater ($k = 1, 2$); γ_{000} is the intercept for a fixed effect of the group average; the following three terms represent random effects for the ith participant, the jth tracing order, the kth rater, respectively; and other three terms denote random interaction effects between the jth tracing and the ith participant, between the kth rater and the ith participant, between the jth tracing and the kth rater; and r_{ijk} is an error term.

The above-mentioned model assumes that the seven included variables are independent and distributed normally with zero means. The total variances can be decomposed into the variance component:

- among participants $\sigma^2_{\text{subject}}$
- between repeated tracings by the same rater σ^2_{order}
- between raters for the same tracing order σ^2_{rater}
- among participants due to the differences in tracing order $\sigma^2_{\text{subject}\times\text{order}}$
- among participants due to the differences in rater $\sigma^2_{\text{subject}\times\text{rater}}$
- between two raters due to the differences in tracing order $\sigma^2_{\text{order}\times\text{rater}}$
- of the residual σ^2_r.

We define the inter-rater reliability of the human amygdala volumetric measurements by manual tracing as:

$$\text{interICC} = \frac{\sigma^2_{\text{subject}}}{\sigma^2_{\text{subject}} + \sigma^2_{\text{rater}} + \sigma^2_{\text{subject}\times\text{rater}} + \sigma^2_r} \tag{2}$$

and the intra-rater reliability of the human amygdala volumetric measurements by manual tracing as:

$$\text{intraICC} = \frac{\sigma^2_{\text{subject}}}{\sigma^2_{\text{subject}} + \sigma^2_{\text{order}} + \sigma^2_{\text{subject}\times\text{order}} + \sigma^2_r} \tag{3}$$

2.3.2. Automatic tracing and visual inspection

Amygdala volumes were estimated using *volBrain* (http://volbrain.upv.es), a fully automated segmentation method that has outperformed other segmentation methods across many brain structures (Manjón and Coupé, 2016). The operational pipeline has been described and evaluated previously (Manjón and Coupé, 2016). Intracranial volume (ICV) and total gray matter volume (GMV) were derived with *volBrain*. Amygdala volumes were also obtained using *FSL*-FIRST (v.6.0.4; http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST). More detailed information about the processing steps of subcortical segmentation by *FSL*-FIRST can be found in Patenaude et al. (2011). In the current study, we segmented the T1 images using *FSL*-FIRST with none boundary correction. Amygdala segmentation labels were saved as binary masks. A voxel count was subsequently used to calculate the amygdalar volumes. Both ICV and GMV were also obtained using *FSL*. *FreeSurfer* segmentation includes the cross-sectional (*CS-FS*) and the longitudinal streams (*LG-FS*). Automatic segmentation and labeling of the human amygdala were also performed using the 'recon-all' pipeline in *CS-FS* and *LG-FS* (ver. 6.0.0; http://surfer.nmr.mgh.harvard.edu). These processing stages have been documented in Fischl et al. (2002) and Reuter et al. (2012). Amygdala volumes provided in **aseg.stats** files were used in the subsequent analysis, and **aseg.mgz** volume files were converted into NIFTI files in native space for visualization. Transformation for segmentation and its inverse transformation to native space for volumetric comparison have been described in Morey et al. (2009). Both ICV and GMV measurements were provided by the *CS-FS* outputs while The GMV measurements were also provided by the LG-FS outputs. Due to *LG-FS*'s possible bias by matching head across all the time points in children, ICV measurements were not obtained by *LG-FS*. All segmentation results were visually inspected to ensure adequate registration. Visual inspection of the traced amygdala volumes in a representative subject using manual and automatic methods is illustrated in Fig. 1.

Table 3
Comparison of automated segmentations to manual tracing.

Wave	Technique	Structure volume (mean cm^3 ±SD)		Comparison of techniques to manual tracing									
				%Volume difference ±SD		%Volume overlap ±SD		%False positive ±SD		%False negative ±SD		Correlation	
		Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
	Manual	1.46 ± 0.18	1.45 ± 0.18										
	volBrain	0.90 ± 0.12***	0.92 ± 0.13***	38.18 ± 7.60	36.48 ± 7.59	68.16 ± 4.92	68.21 ± 4.86	10.28 ± 5.17	11.65 ± 5.95	44.73 ± 5.73	44.13 ± 5.65	0.61***	0.62***
Wave 1	CS-FS	1.65 ± 0.21***	1.66 ± 0.25***	16.96 ± 11.50	18.78 ± 13.99	78.32 ± 3.65	78.78 ± 3.33	26.76 ± 5.61	26.69 ± 6.13	15.36 ± 5.06	14.22 ± 4.81	0.62***	0.59***
	LG-FS	1.72 ± 0.22***	1.78 ± 0.26***	22.85 ± 13.16	28.08 ± 15.79	74.88 ± 3.84	75.97 ± 3.41	31.56 ± 5.66	31.79 ± 5.98	16.80 ± 5.19	13.60 ± 4.40	0.61***	0.57***
	FSL	1.50 ± 0.30	1.48 ± 0.33	19.59 ± 13.43	20.95 ± 14.46	79.19 ± 4.94	77.05 ± 7.07	20.49 ± 8.89	22.45 ± 8.38	19.02 ± 11.21	21.03 ± 13.63	0.09	0.08
	Manual	1.48 ± 0.17	1.48 ± 0.18										
	volBrain	0.92 ± 0.11***	0.93 ± 0.12***	37.89 ± 6.53	37.25 ± 7.01	67.78 ± 4.06	68.23 ± 4.67	11.17 ± 4.81	11.05 ± 5.27	44.99 ± 4.81	44.38 ± 5.41	0.63***	0.61***
Wave 2	CS-FS	1.65 ± 0.22***	1.63 ± 0.23***	15.93 ± 10.67	13.94 ± 10.62	77.69 ± 3.46	79.11 ± 3.38	26.78 ± 4.95	24.75 ± 5.40	16.71 ± 6.09	15.89 ± 5.34	0.54***	0.66***
	LG-FS	1.74 ± 0.22***	1.77 ± 0.23***	21.64 ± 13.17	23.88 ± 12.31	75.40 ± 3.26	76.78 ± 3.34	30.83 ± 5.13	30.18 ± 4.89	16.59 ± 5.33	14.24 ± 5.00	0.53***	0.64***
	FSL	1.55 ± 0.32*	1.52 ± 0.30	20.48 ± 16.28	18.21 ± 12.87	79.03 ± 5.20	77.28 ± 5.97	21.24 ± 9.15	22.58 ± 7.91	18.48 ± 11.20	20.96 ± 11.92	−0.03	0.14
	Manual	1.49 ± 0.21	1.51 ± 0.21										
	volBrain	0.92 ± 0.15***	0.94 ± 0.16***	38.44 ± 7.71	37.89 ± 8.09	67.54 ± 4.87	67.34 ± 5.33	10.81 ± 5.13	11.59 ± 5.13	45.32 ± 5.79	45.27 ± 6.34	0.67***	0.70***
Wave 3	CS-FS	1.63 ± 0.23***	1.68 ± 0.25***	14.35 ± 10.85	16.04 ± 13.23	76.90 ± 4.88	77.25 ± 4.15	26.91 ± 5.70	27.06 ± 6.44	18.32 ± 7.21	17.24 ± 5.45	0.66***	0.59***
	LG-FS	1.72 ± 0.24***	1.80 ± 0.26***	20.71 ± 13.60	26.29 ± 21.86	74.52 ± 4.28	74.78 ± 8.19	31.25 ± 5.37	32.03 ± 8.29	18.05 ± 6.87	16.22 ± 9.39	0.61***	0.59***
	FSL	1.58 ± 0.33*	1.52 ± 0.30	21.50 ± 17.89	17.79 ± 13.33	79.52 ± 4.26	77.68 ± 4.71	21.42 ± 9.59	21.21 ± 7.50	17.65 ± 10.81	21.37 ± 11.35	−0.05	0.15

Note: Means, standard deviations, and the statistical significance of the two paired sample t-test between manual and automated segmentations as well as summary of automated segmentation performance, mean percentage of volume difference, percentage of volume overlap, percentage of false positive and Pearson's correlations between automated and manual segmentations. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

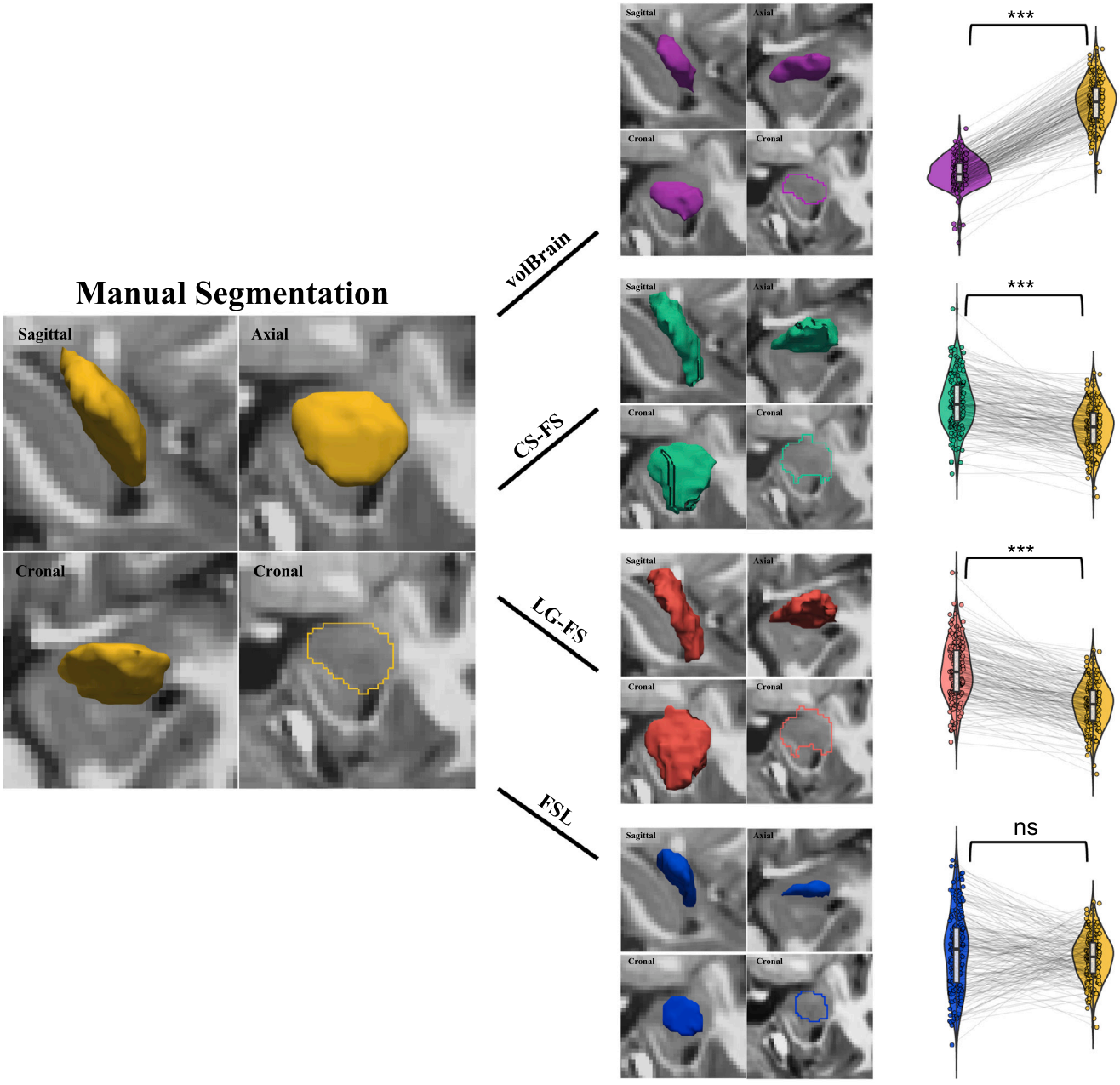


Fig. 1. The amygdala structures extracted using five different techniques. A sample participant's segmented amygdala by manual tracing (Yellow), *volBrain* (Purple), *CS-FS* (Green), *LG-FS* (Red) and *FSL* (Blue). Scatter Violin plots present the paired changes of the traced left amygdala volumes between manual tracing and *volBrain*, *CS-FS*, *LG-FS* and *FSL* for the first-wave devCCNP-SWU samples. $* p < 0.05$; $** p < 0.01$; $*** p < 0.001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.4. Accuracy assessments on automatic segmentation

QZ manually traced all the 427 amygdala of the devCCNP-SWU samples, which served as the reference volumes (i.e., gold standard) for the subsequent analyses. We validated the accuracy of automatic segmentation separately for each of the three waves of the samples. For each wave, we performed paired t-tests on traced volumes between the automatic and manual methods. We quantified volume difference between the automatic and manual tracing as Eq. (4). A greater volume difference indicates increased discrepancy relative to the manually segmented amygdala volumes. To examine systematic changes of the traced volumes, we tested the Pearson's correlation of traced volumes between the automatic and manual methods across individual subjects. A strong correlation ($R \geq 0.8$) is taken to indicate good consistency on the individual differences in amygdala volumes between the manual and automatic methods. We further calculated the spatially overlapping volumes, the false positive rate and the false negative rate to quantitatively measure the degree of correct or incorrect estimation of the automatic methods. These metrics are defined as:

- percentage of volume difference

$$D(V_A, V_M) = \frac{|V_A - V_M|}{V_M} \times 100\% \quad (4)$$

- percentage of spatial overlap

$$P(V_A, V_M) = \frac{V_A \cap V_M}{0.5(V_A + V_M)} \times 100\% \quad (5)$$

- false-positive rate

$$FP(V_A, V_M) = \frac{V_A - V_A \cap V_M}{V_A} \times 100\% \quad (6)$$

- false-negative rate

$$FN(V_A, V_M) = \frac{V_M - V_A \cap V_M}{V_M} \times 100\% \quad (7)$$

In these equations, V_A is the volume measured automatically and V_M is that measured manually (the reference, i.e., the gold standard). The maximum $P(V_A, V_M)$ value is 100%, reflecting identical tracing between manual and automatic method while smaller values indicated less perfect spatial overlaps (Morey et al., 2009), implying the worse performance of the automatic tracing. The minimum $FP(V_A, V_M)$ value is 0, reflecting identical tracing between manual and automatic method while larger values indicate higher error rates of automatic segmentation, i.e., the inclusion of larger proportions of non-amygdalar structure(s). The minimum $FN(V_A, V_M)$ value is 0, reflecting identical tracing between manual and automatic segmentation while higher values indicate more error rates of automated protocol, i.e., the exclusion of larger proportions of amygdalar structure(s).

To further investigate how the accuracy of the automatic tracing methods varies with amygdala sizes, we employed a generalized additive mixed model (GAMM) to model the size effect of amygdala on the automatic tracing accuracy. In addition, young participants are more likely to move during scan acquisition, leading to worse scan quality and motion artifacts, which may affect the precise differentiation of structures by automated techniques. To exclude the effects of image quality on segmentation accuracy, we included motion and scan quality as covariates for the regression models. Specifically, the motion metric is the coefficient of joint variation (CJV) as an objective function for the optimization of intensity non-uniformity correction algorithms (Ganzetti et al., 2016) while the image quality is quantified by signal-to-noise ratio (SNR). Lower CJV and higher SNR values indicate better image quality (Welvaert and Rosseel, 2013). Specifically, we plotted the spatial overlap (the overlap percentage P) between automatic and manual segmentation as a function of the reference (i.e., the manually traced) volumes, while controlling CJV and SNR. Unlike the common parametric linear models (Herting et al., 2018), GAMM does not require a-priori knowledge of the relationship between the response and predictors, which enables more flexible and efficient estimation of changing patterns (Mills and Tamnes, 2014; Wood, 2017). In addition, GAMMs are well suited for the repeated measurements (e.g., our accelerated longitudinal samples from developing brains), as they account for both within-subject dependency and developmental differences among participants at the time of study enrollment (Alexander-Bloch et al., 2014; Harezlak et al., 2005). Such a GAMM was implemented using the following formula in *R* language with the **mgcv** package:

$$P(V_A, V_M) \sim s(V_M) + \text{CJV} + \text{SNR} + (1|\text{subject}) \quad (8)$$

where the $s()$ is a smoothing function with a fixed degree of freedom and cubic B-splines, whose number of knots is set at 5 (determined to be optimal for our data). This was set to be sufficiently large to have adequate degrees of freedom across both spline terms from fits of the model to the amygdala volume, but sufficiently small to maintain reasonable computational efficiency. The CJV and SNR were included as fixed-effect terms in the GAMM regression.

2.5. Modeling growth curves of human amygdala development

To fully model method-related differences in the growth curves of human amygdala volumes, we employed the following GAMM to examine age-related changes of the human amygdala by including the tracing method and its interaction with age as variables of interests:

$$V \sim s(age) + \text{method} + s(age, by = \text{method}) + \text{sex} + (1|\text{subject}) \quad (9)$$

where V represents the amygdala volume and $s()$ is a smoothing function, with a fixed degree of freedom and cubic B-splines (the number of knots = 5). Tracing method was entered as an ordinal factor (manual = 0, automatic = 1). The method term reflects the method differences in the intercept (i.e., the main effect of method). The sex term reflects the sex differences in the intercept (i.e., the main effect of sex). The first smoothing term models the slope of age for manual tracing, and the second smoothing term models the difference in the age-related slope between methods (i.e., *age*×method interaction). The p value associated with this term is the basis of statistical inference regarding methodological differences in developmental trajectories of bilateral amygdala volume.

To more specifically understand differences in age trajectories between methods, a set of GAMMs (see the Eq. (10)) were proposed to detect age-related changes revealed by each method separately:

$$\begin{aligned} V &\sim age + (1|\text{subject}) \\ V &\sim age + \text{sex} + (1|\text{subject}) \\ V &\sim s(age) + (1|\text{subject}) \\ V &\sim s(age) + \text{sex} + (1|\text{subject}) \\ V &\sim s(age) + \text{sex} + s(age, by = \text{sex}) + (1|\text{subject}) \end{aligned} \quad (10)$$

The first GAMM models the traced volume as a fixed linear age effect. As previous studies have consistently shown larger brain regions in males than in females (Herting et al., 2018), we established the second GAMM model with sex as a fixed term to assess the sex difference in the trajectory intercept. The third GAMM models the traced volume as a smoothing function of age. The fourth model is established by including sex as a fixed-effect term in the third model, as well as the fifth GAMM model including *age*×sex interaction to test the sex differences in the trajectory slope. The Akaike Information Criterion (AIC) was used to determine which model had the best fit. All fit models were tested against a null age effect model. The model chosen as the best fit model had to have the lowest AIC value and be significantly different from null. The data analyses and visualization were performed using the **mgcv** (Wood, 2017) and **ggplot2** (Wickham, 2016) packages in R (R Core Team, 2014), respectively.

We also tested growth curves of the human amygdala by accounting for global brain features in the GAMMs. The volumes of subcortical structures are known to be related to brain size (Brown et al., 2014; Brain Development Cooperative Group, 2012; Uematsu et al., 2012). Accordingly, we included ICV as a covariate for regression control to enable the removal of individual variability that can be explained by brain size (Narvacan et al., 2017; Sawiak et al., 2018; Herting et al., 2014). Researchers have also demonstrated that the size of the amygdala often scales with the GMV (Van Petten, 2004; Rice et al., 2014). We thus accounted for brain size by controlling for the GMV in the GAMMs. We performed the analysis with *CS-FS*, *LG-FS*, *volBrain* and *FSL* derived GMV measurements and *CS-FS*, *volBrain* and *FSL* derived ICV measurements. To better understand the differential effects when controlling for ICV and GMV, we modeled and plotted the measurement bias as a function of either GMV or ICV. The measurement bias were quantified by the spatial overlap and false positive rate of automated segmentation with manual tracing. The growth curves were delineated for both GMV and ICV to help understanding their differential effects when controlling them for modeling amygdala's growth.

3. Results

3.1. Measurement reliability of manually traced human amygdala

We reported almost perfect reliability of the human amygdala volumes measured by the manual tracing protocol. Specifically, as in Table 2, both intra-rater and inter-rater reliability of the volumes for the manually traced amygdala were achieved. Inter-rater ICCs were around 0.88 with 95%CI = [0.80, 0.96] for the left amygdala, and 0.89

with 95%CI = [0.83, 0.95] for the right amygdala. Intra-rater ICCs were also almost perfect: 0.91 with 95%CI = [0.82, 0.96] for rater ZQZ while 0.95 with 95%CI = [0.89, 0.97] for rater QZ. These results confirmed that the raters' manual tracings could be used as the gold standard or the reference for comparisons with automatic segmentation.

3.2. Measurement accuracy of automatically traced human amygdala

For the first-wave samples, one-way analysis of variance with repeated measures indicated significant differences in volumes of human amygdala across the five segmentation methods (left amygdala: $F = 413.90, p < 0.001$; right amygdala: $F = 347.70, p < 0.001$). Our post-hoc paired comparisons between the automatic and manual methods revealed that volumes obtained with *CS-FS* and *LG-FS* were both significantly larger than those obtained by manual tracing (*CS-FS*: left amygdala: $t = 15.45, p < 0.001$, right amygdala: $t = 14.51, p < 0.001$; *LG-FS*: left amygdala: $t = 19.94, p < 0.001$, right amygdala: $t = 21.07, p < 0.001$, respectively), which in turn were larger than those obtained with *volBrain* (left amygdala: $t = 53.32, p < 0.001$; right amygdala: $t = 50.09, p < 0.001$). The amygdala volumes obtained by manual tracing were not significantly different with those obtained by *FSL* (left amygdala: $t = 1.75, p = 0.081$; right amygdala: $t = 1.36, p < 0.175$). These findings (Fig. 1) are reproducible for the second and third waves of samples with the exceptions of the left amygdala volumes obtained with *FSL* were significantly larger than those obtained with manual tracing (wave-2: $t = 2.27, p = 0.025$; wave-3: $t = 2.10, p = 0.038$; see Supplementary Figure S1 and S2).

As depicted in Fig. 2, paired two-sample t-tests revealed that the spatial overlap for the left amygdala were significantly higher for *FSL* than *CS-FS* ($t = 2.10, p = 0.037$), while the spatial overlap for the right amygdala were significantly higher for *CS-FS* than *FSL* ($t = 3.14, p = 0.002$). The comparisons between the spatial overlap of *CS-FS* and *FSL* were inconsistent for the three waves. Both *CS-FS* and *FSL* methods showed higher spatial overlap with manual tracing than *LG-FS* (left amygdala: $t = 19.23, p < 0.001$; right amygdala: $t = 17.69, p < 0.001$; left amygdala: $t = 9.91, p < 0.001$; right amygdala: $t = 1.98, p = 0.05$). The *LG-FS* had higher percentages of spatial overlap than *volBrain* with the manual tracing for the first-wave data (left amygdala: $t = 14.66, p < 0.001$; right amygdala: $t = 19.47, p < 0.001$). The false positive rate of amygdala were significantly different between all methods, with the *volBrain* showing the lowest value, followed by *FSL*, *CS-FS* and *LG-FS*. These findings are reproducible for the second-wave and the third-wave samples (see Supplementary Figures S1, S2). The false-negative rates were significantly higher for *volBrain* than for *FSL*, *CS-FS* and *LG-FS* segmentation of amygdala (all $ps < 0.05$). The *FSL*, *CS-FS* and *LG-FS* showed comparable false negative rates, which resulting in inconsistency comparison results for three waves of samples (Fig. 2, Figures S1, S2). Both the left and right amygdala volumes obtained with the *CS-FS*, *LG-FS* and *volBrain* methods only showed moderate Pearson's correlations with those obtained by the manual tracing although statistically significant ($Rs = 0.57 - 0.62, ps < 0.001$; Table 3), but did not exceed 0.8. Correlations between manual and *FSL* were obviously weaker and did not reach statistical significance (left amygdala: $r = 0.09, p = 0.207$; right amygdala: $r = 0.08, p = 0.310$). These indicated that the individual differences in volume measured by the automatic methods were not fully consistent with those measured by the manual tracing method.

The GAMM-based regression showed that the accuracy of *volBrain* segmentation (i.e., the percentage of spatial overlap with manual tracing) increased with the amygdala size before reaching a stable accuracy with a larger volume of the amygdala (Fig. 3; see details on the parameters for models in Supplementary **Table S1**). For *CS-FS* and *LG-FS*, the segmentation accuracy displayed a linearly significant increase pattern with the left amygdala size while a two-stage (first increase and then remain stable) pattern with the right amygdala size. The *FSL* increased and then followed by plateau with the left amygdala size while consistently remained stable with right amygdala size. In most cases of the automatic segmentation methods, a smaller amygdala structure is associated with worse segmentation accuracy after controlling image quality. These results indicated that neuroanatomical features can possibly bias the accuracy of automatic segmentation in a systematic way. In addition, the effects of image quality and motion are detectable for both *volBrain* and *FreeSurfer* segmentation accuracy while not *FSL*. The effect of more movements (higher CJV) on the segmentation accuracy of amygdala with *CS-FS* was significant in children during imaging while the effect of worse quality (lower SNR) on the segmentation accuracy with both *volBrain* and *FreeSurfer* was also significant (see in Supplementary Table S1).

3.3. Growth curves of human amygdala volume

The unified GAMM method, which includes age and interactions terms indicated that the age effects on the human amygdala were not consistent across the automatic tracing methods (Table 4). Specifically, these models reproduced the results of measurement accuracy for *volBrain*, *FreeSurfer* and *FSL* reported in the previous section. The *volBrain* produced amygdala's age-related changes highly similar to that of manual tracing, i.e., no age×method interactions (all $ps > 0.05$). In contrast, the age-related amygdala changes showed discrepancies between *FreeSurfer* (including both *CS-FS* and *LG-FS*) and the manual tracing more overt than *volBrain* versus manual tracing method-differences, although no statistically significant age×method interaction was detected (all $ps > 0.05$). This led to a much lower explained variance using the GAMMs with *CS-FS* and *LG-FS* compared to that by the GAMMs with *volBrain* (*CS-FS*: left amygdala: 28% versus 80%, right amygdala: 28% versus 77%; *LG-FS*: left amygdala: 39% versus 80%, right amygdala: 33% versus 77%, respectively). In contrast, the statistically-significant age×method (*FSL* versus manual tracing) interaction was detectable (left amygdala: $p < 0.001$; right amygdala: $p = 0.001$). This indicated a significant difference in the growth rate of the bilateral amygdala volume between the *FSL* and manual segmentation.

The post-hoc method-wise GAMMs further revealed the growth patterns of the human amygdala as well as their sex differences. All best-fitting growth curves for each method determined by AIC were the smoothing-age models, which were significantly different from the null model on age effect. For all methods except *FSL*, the best models were determined by AIC as the fourth model, which included sex as a fixed effect and age adjusted for a smoothing spline function (Table 5), indicating no need for an interaction between age and sex. This model revealed bigger amygdalae in boys than in girls, but their growth rates did not differ by sex. Specifically, as shown in Fig. 4, the growth curve patterns were parallel in girls and boys for both manual and automatic tracing methods although boys demonstrated larger volumes of their amygdalae than girls across the entire school age range (6–18 years old). As the reference standard, the manual tracing method revealed that the human amygdala (both left and right) exhibited linear growth during the school-age years in both boys and girls (left amygdala: $p = 0.003$; right amygdala: $p = 0.001$, respectively). The *volBrain* traced left amygdala yielded less linear growth ($p < 0.001$), and traced right amygdala yielded growth curves very similar to that established by the manual tracing method ($p = 0.001$). *CS-FS* tracing method produced less linear and flatter curves and not statistically significant, except for a marginal significant growth curve in the right amygdala ($p = 0.066$). This growth curve had an inverted U shape: increasing during childhood and early adolescence, and then decreasing in late adolescence (the peak age around 14.18 years old). *LG-FS* produced similar shape with those of *CS-FS*, all exhibiting somehow nonlinearity although its degree of nonlinearity is left–right flipped between the two *FS* segmentation methods. In addition, *LG-FS* detected statistical significant increases with age (left amygdala: $p = 0.034$; right amygdala: $p = 0.022$, respectively). For *FSL* method, the best model was the third model, which only included the smooth age effect. This model revealed

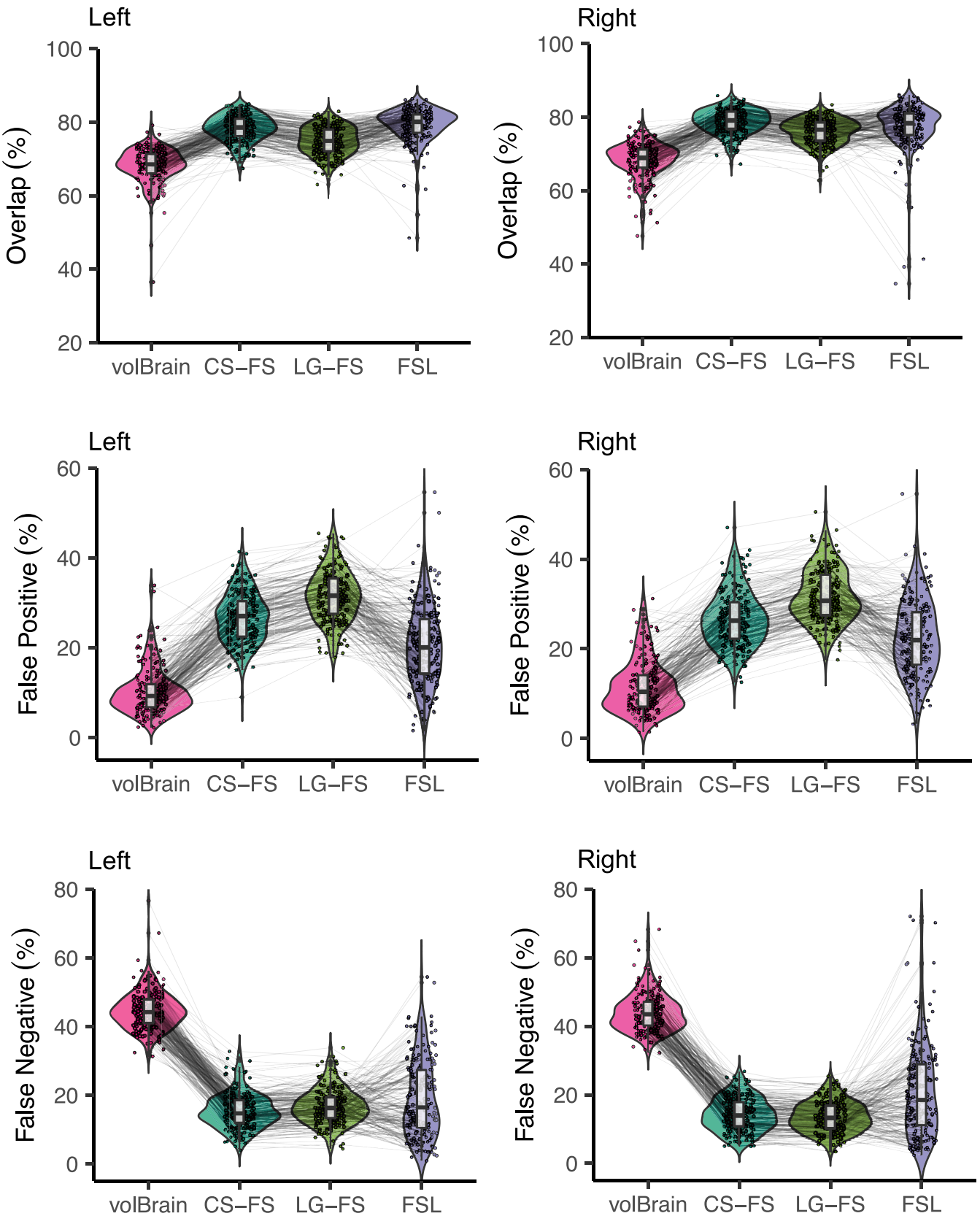


Fig. 2. Spatial overlap, false positive rate and false negative rate for segmentation using *volBrain*, *CS-FS*, *LG-FS* and *FSL* compared to the manual 'gold standard' in the devCCNP-SWU wave-1 samples.

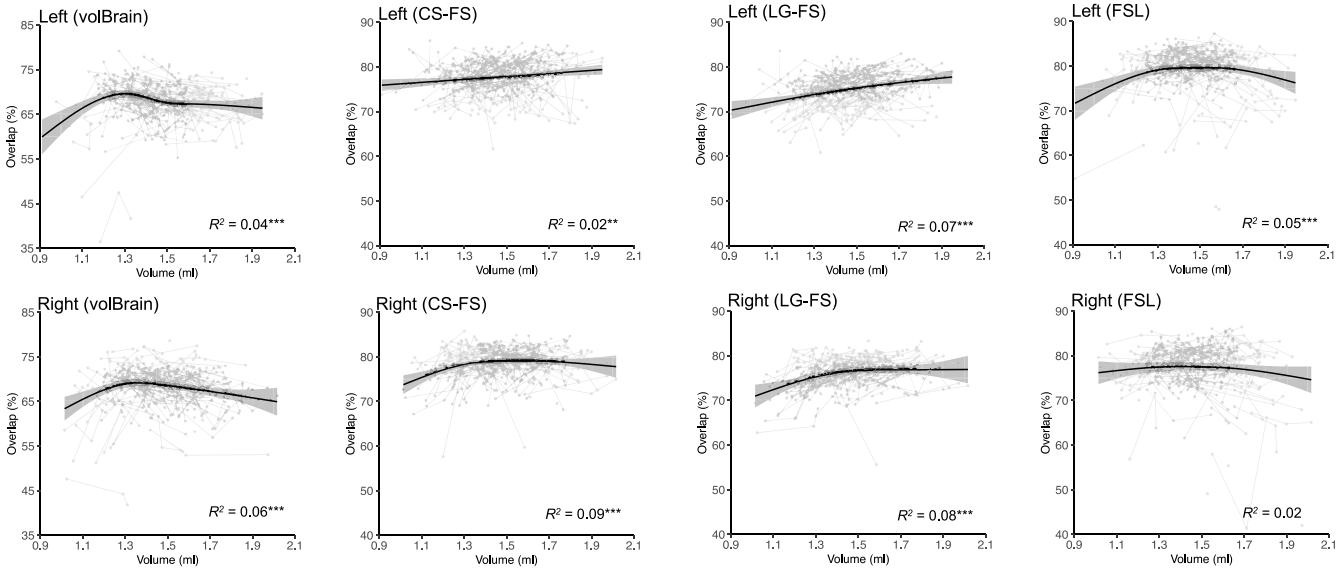


Fig. 3. Percentage of spatial overlap of automatic methods as function of the amygdala volume. $* p < 0.05$; $** p < 0.01$; $*** p < 0.001$.

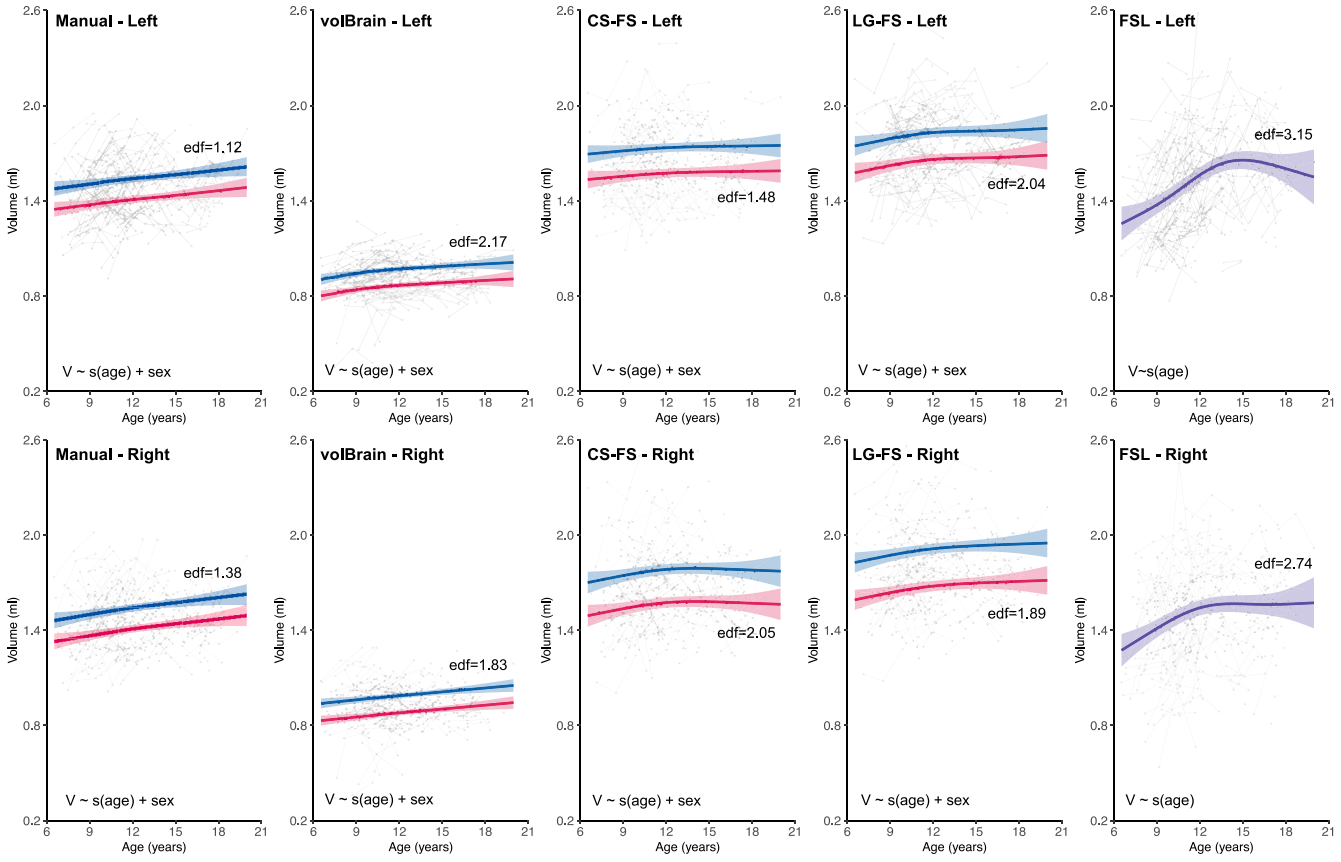


Fig. 4. Volumetric growth curves for human amygdala traced by manual tracing, *volBrain*, CS-FS, LG-FS and *FSL*. The blue color indicates trajectories for boys while the red color indicates trajectories for girls while the purple color for each sex plotted together. The trajectories are surrounded by shaded 95% confidence intervals. Note that boys and girls showed very similar developmental trajectories with no significant age-by-sex interactions, although boys had significantly larger amygdala volumes across the school ages (all *ps* < 0.001), except *FSL* showing the same trajectories for boys and girls. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the same pattern of trajectories for girls with boys while the bilateral amygdala exhibited two-phase growth patterns significantly different from that established by the manual tracing: robust volume increase during childhood and gradual decrease in late adolescence in left

Table 4
GAMM estimates on age, method, and age*method effects for bilateral amygdala.

Left Amygdala					Right Amygdala				
Manual vs. volBrain									
Intercept	**Estimate**	**SE**	***t***	***p* - value**	**Intercept**	**Estimate**	**SE**	***t***	***p* - value**
Method: volBrain	−0.40	0.00	−81.68	**0.000**	Method: volBrain	−0.38	0.00	−77.35	**0.000**
Sex	0.12	0.02	7.10	**0.000**	Sex	0.12	0.02	6.88	**0.000**
Slope	**edf**	**Ref.df**	**F**	**p - value**	**Slope**	**edf**	**Ref.df**	**F**	**p - value**
s (age)	2.98	2.98	6.43	**0.000**	s (age)	2.44	2.44	8.47	**0.000**
s (age): volBrain	1.00	1.00	0.80	0.371	s (age): volBrain	1.53	1.53	0.41	0.706
R^2 = 0.80					R^2 = 0.77				
Manual vs. CS-FS									
Intercept	**Estimate**	**SE**	***t***	***p* - value**	**Intercept**	**Estimate**	**SE**	***t***	***p* - value**
Method:CS-FS	0.12	0.01	19.80	**0.000**	Method: CS-FS	0.13	0.01	19.81	**0.000**
Sex	0.14	0.02	6.78	**0.000**	Sex	0.17	0.02	7.41	**0.000**
Slope	**edf**	**Ref.df**	**F**	**p - value**	**Slope**	**edf**	**Ref.df**	**F**	**p - value**
s (age)	2.23	2.23	3.36	**0.041**	s (age)	1.76	1.76	5.70	**0.003**
s (age): CS-FS	1.00	1.00	0.14	0.710	s (age): CS-FS	1.84	1.84	1.45	0.154
R^2 = 0.28					R^2 = 0.28				
Manual vs. LG-FS									
Intercept	**Estimate**	**SE**	***t***	***p* - value**	**Intercept**	**Estimate**	**SE**	***t***	***p* - value**
Method: LG-FS	0.18	0.01	28.86	**0.000**	Method: LG-FS	0.22	0.01	33.00	**0.000**
Sex	0.15	0.02	6.85	**0.000**	Sex	0.18	0.02	7.84	**0.000**
Slope	**edf**	**Ref.df**	**F**	**p - value**	**Slope**	**edf**	**Ref.df**	**F**	**p - value**
s (age)	3.11	3.11	3.86	**0.010**	s (age)	1.28	1.28	9.09	**0.001**
s (age): LG-FS	1.00	1.00	0.02	0.896	s (age): LG-FS	1.90	1.90	1.30	0.207
R^2 = 0.39					R^2 = 0.33				
Manual vs. FSL									
Intercept	**Estimate**	**SE**	***t***	***p* - value**	**Intercept**	**Estimate**	**SE**	***t***	***p* - value**
Method:FSL	0.04	0.01	4.56	**0.000**	Method: FSL	0.02	0.01	1.88	0.060
Sex	0.09	0.03	3.53	**0.000**	Sex	0.10	0.02	4.83	**0.000**
Slope	**edf**	**Ref.df**	**F**	**p - value**	**Slope**	**edf**	**Ref.df**	**F**	**p - value**
s (age)	1.00	1.00	1.50	0.222	s (age)	1.00	1.00	2.45	0.118
s (age): FSL	2.89	2.89	15.64	**0.000**	s (age): FSL	2.18	2.18	2.21	0.135
R^2 = 0.12					R^2 = 0.05				

Note: Smooth function (edf) as well as degrees of freedom (Ref.df) and *F*-statistic and associated *p*-value for age (***bold*** highlights $p<.05$).

Table 5
GAMM estimates for bilateral amygdala across both sexes in manual tracing, volBrain, FSL, CS-FS, and LG-FS segmentation separately.

	Hemisphere	Best model fit	R^2(adjusted)	Sex				Age spline			
				Estimate	SE	*t*	*p*-value	edf	Red.df	*F*	*p*-value
Manual	Left amygdala	s(age) + sex	0.14	0.13	0.02	6.07	**0.000**	1.12	1.12	8.32	**0.003**
	Right amygdala	s(age) + sex	0.12	0.12	0.02	5.86	**0.000**	1.38	1.38	8.47	**0.001**
volBrain	Left amygdala	s(age) + sex	0.19	0.10	0.01	7.15	**0.000**	2.17	2.17	8.07	**0.000**
	Right amygdala	s(age) + sex	0.17	0.11	0.02	6.84	**0.000**	1.83	1.83	7.34	**0.001**
CS-FS	Left amygdala	s(age) + sex	0.15	0.16	0.03	6.08	**0.000**	1.48	1.48	0.69	0.317
	Right amygdala	s(age) + sex	0.17	0.21	0.03	7.15	**0.000**	2.05	2.05	2.77	0.066
LG-FS	Left amygdala	s(age) + sex	0.16	0.17	0.03	6.13	**0.000**	2.04	2.04	3.52	**0.034**
	Right amygdala	s(age) + sex	0.20	0.24	0.03	7.81	**0.000**	1.89	1.89	3.42	**0.022**
FSL	Left amygdala	s(age)	0.11					3.15	3.15	19.48	**0.000**
	Right amygdala	s(age)	0.06					2.74	2.74	11.13	**0.000**

Note: Smooth function (edf) as well as degrees of freedom (Ref.df) and *F*-statistic and associated *p*-value (***bold*** highlights $p < .05$) for age.

amygdala ($p < 0.001$; the peak age around 15.20 years); robust volume increase during childhood and a phase of plateau over adolescence in right amygdala ($p < 0.001$; the peak age around 16.03 years).

Correction for GMV abolished the significant sex differences of trajectory patterns derived by *volBrain* across the entire age range (see details on the parameters for best-fitting models in Supplementary **Table S2**). The growth patterns derived by the manual tracing method after controlling for either *volBrain*-estimated GMV or *FreeSurfer*-estimated GMV or *FSL*-estimated GMV remain consistent with those without the GMV corrections (Fig. 5). After the GMV-based correction, the growth patterns derived by *volBrain*, *CS-FS* and *LG-FS* showed almost identical shapes to those obtained using the manual tracing method (Fig. 6). This correction highly increased the reproducibility of the human amygdala growth curves across *volBrain*, *CS-FS* and *LG-FS*. The shape discrepancies between manual tracing and *FSL* derived growth curves after the GMV-based correction did not resolved. As shown in Fig. 5, the growth patterns derived by the manual tracing method remained consistent with those without ICV corrections, but with much less statistical power: controlling for *volBrain*-estimated ICV or *FSL* estimated ICV led to much less significant age-related changes while controlling for *CS-FS*-estimated ICV led to no significant age-related changes (see Supplementary Table S3). Correction for ICV reduced the reproducibility of the human amygdala growth curves across the three automatic methods (Fig. 6). The significant positive linear association with age remained for *volBrain* traced amygdala with less statistical power, even after controlling for the ICV (Fig. 6). However, correction for ICV changed the *CS-FS*-derived growth curves of the amygdala volume from nonlinear (not significant) to linear decrease (significant) patterns (Fig. 6). The growth patterns derived by the *FSL* were still

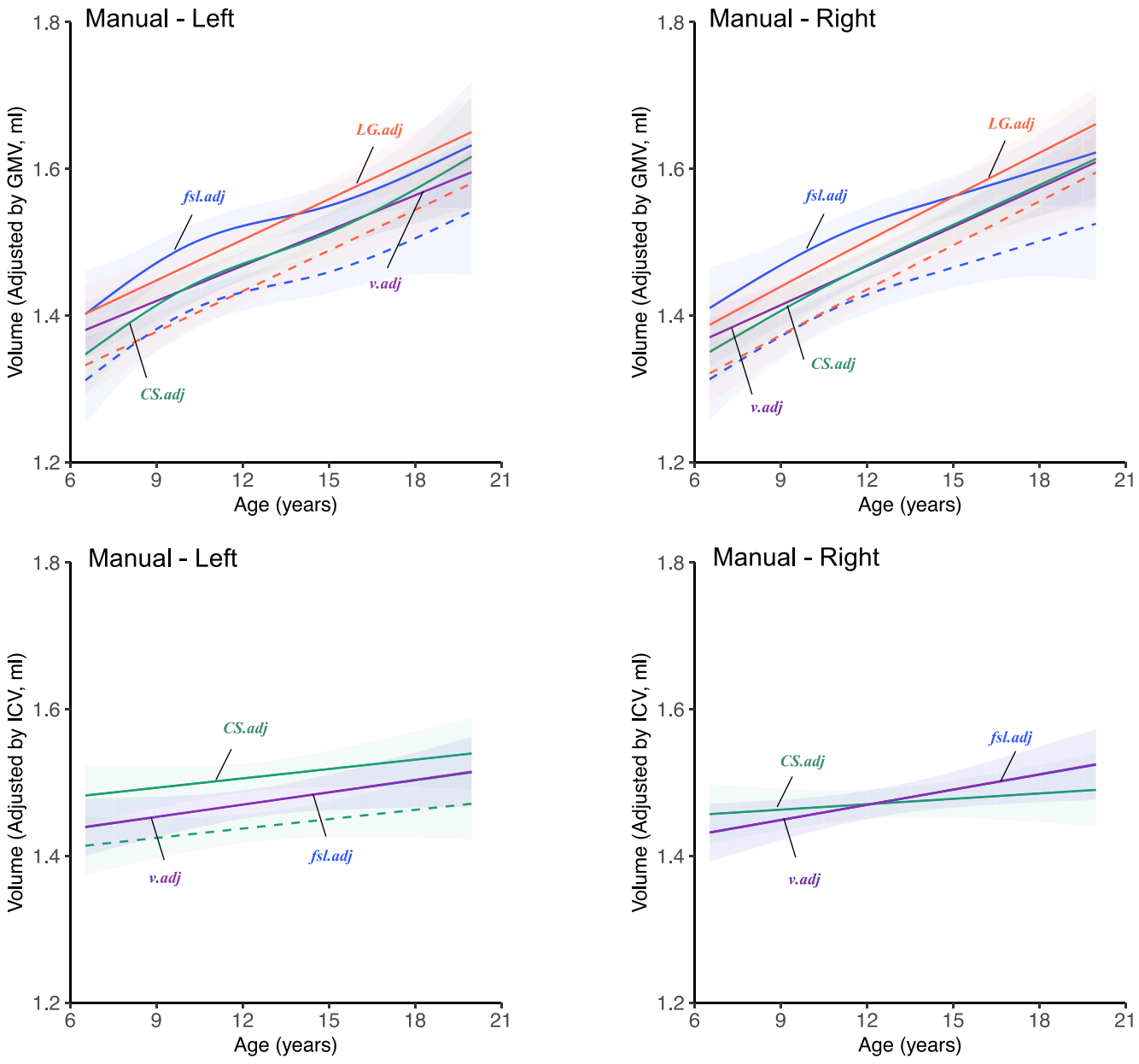


Fig. 5. Growth curves of manually traced volume for human amygdala adjusted by GMV and ICV. Amygdala volumes are adjusted by GMV in different ways: CS.adj, adjusted by *CS-FS* produced GMV; LG.adj, adjusted by *LG-FS* produced GMV; v.adj, adjusted by *volBrain* produced GMV; fsl.adj, adjusted by FSL produced GMV. Amygdala volumes are adjusted by ICV in different ways (B): CS.adj, adjusted by *CS-FS* produced ICV; v.adj, adjusted by *volBrain* produced ICV; fsl.adj, adjusted by *FSL* produced ICV. solid line = male, dotted line = female. The trajectories are surrounded by shaded 95% confidence intervals.

inconsistent with manual tracing without ICV corrections, showing nonlinear increases (Fig. 6).

4. Discussion

This study evaluated the performance of segmentation of the amygdala using the automatic software *volBrain*, the cross-sectional and longitudinal pipeline of *Freesurfer* and *FSL* compared to manual tracing in a longitudinal developmental sample. Importantly, we also explored how the segmentation differences could impact the growth curve modeling of the amygdala development. The findings indicated systematic differences in tracing performance across the three methods. *CS-FS* overestimated the volumes with more spatial overlapping with the manual tracing method, but had higher false-positive rates. *LG-FS* also segmented larger amygdalae than the manual method and showed smaller spatial overlaps and higher false-positive with the manual tracing than *CS-FS* segmentation. In contrast, *volBrain* tended to underestimate the volumes with less spatial overlap with the manual tracing method, but had lower false-positive rates. *FSL* estimated the volumes with more spatial overlapping but had inconsistency with the manual tracing. We noted that the tracing accuracy of automatic methods was worse for smaller amygdalae. Furthermore, the growth curves of the amygdala volume estimated by different methods were inconsistent. These discrepancies indicated the importance to evaluate the segmentation performance across methods, especially in a developmental sample. This study presented manual tracing of the amygdalae in a large-scale longitudinal sample and presented a systematic investigation of the method-wise variability of the growth curves of the human amygdala across school age. This variability of growth patterns

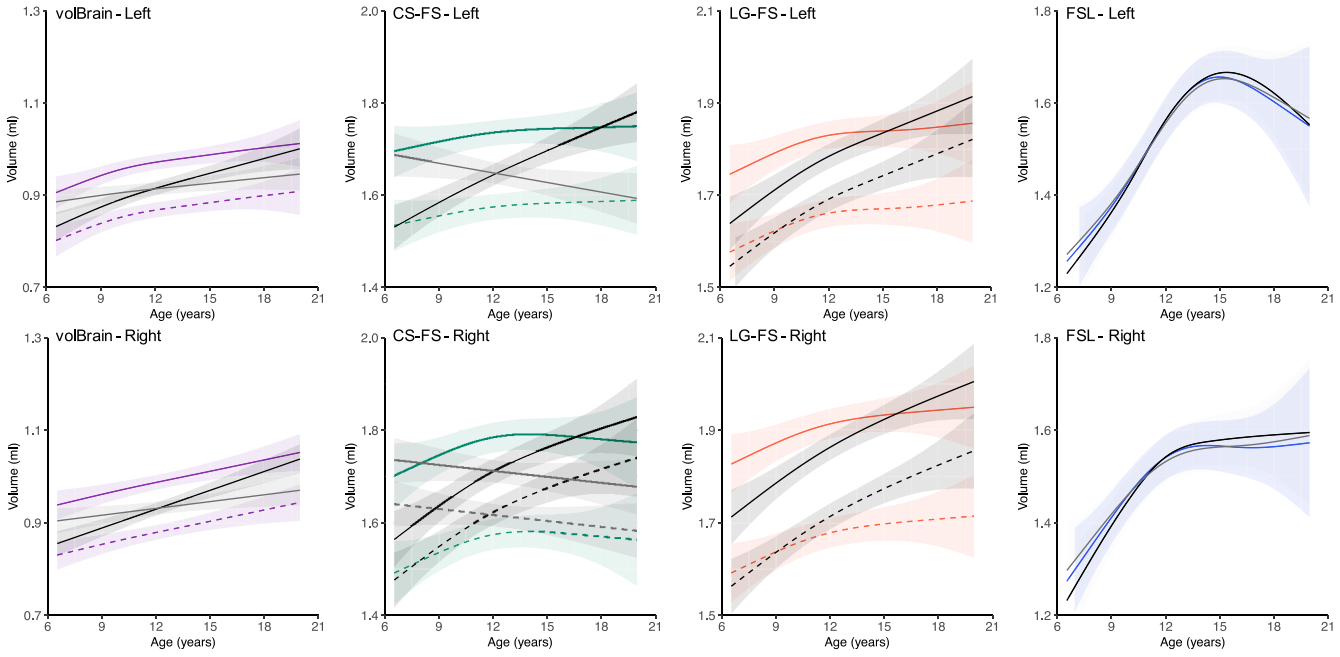


Fig. 6. Growth curves of automatically segmented volume for human amygdala adjusted by GMV (black line) and ICV (gray line). solid line = male, dotted line = female. The trajectories are surrounded by shaded 95% confidence intervals.

of amygdala volume derived from *volBrain* and *FreeSurfer* could be normalized by adjusting for the total gray matter volume, but not adjusting for intracranial volume. The manual tracing method revealed linear growth of the amygdala in both boys and girls throughout the school-aged years, which is valuable to provide a growth norm for pediatric studies in the future.

The measurement accuracy of the amygdala volume varied across the automatic methods. *FreeSurfer* overestimated amygdala volumes (13%–28%), and this overestimation has been observed in previous studies of the amygdala volume measurement by *Freesurfer* (Morey et al., 2009; Schoemaker et al., 2016). It is likely due to the greater variability in the definition of the amygdala boundary and liberal inclusion of voxels near this boundary (Morey et al., 2009; Schoemaker et al., 2016). The degree of overestimation observed here was greater than that reported for adults (7 – 9%) (Morey et al., 2009), but less than that reported for children aged 6–11 years (93%–100%) (Schoemaker et al., 2016). In addition, *FSL* had comparable volume estimates with manual tracing or had slightly higher estimates than the manual method. Although the *FSL* did not excessively overestimate and underestimate amygdala volume, it showed high absolute difference percentage (17%–21%) with manual tracing. The degree of volume difference observed here was also greater than that reported for adults (3–6%) (Morey et al., 2009), but less than that reported for children aged 6–11 years (40%–50%) (Schoemaker et al., 2016). Schoemaker et al. (2016) suggested it might be caused by using a standard brain template derived from adults. This may introduce greater bias when applied to a pediatric sample, in which amygdala sizes and shapes differ from adults. Another possible reason for this volume difference might be artifacts caused by more movements in children during imaging, causing a less precise differentiation and classification of amygdala structures by *FreeSurfer*. In contrast, *volBrain* underestimated the amygdala volume (35%–37%) compared to the manual tracing. The underestimation may reflect the stringent inclusion of the amygdala during the segmentation by *volBrain*. This underestimation has been also observed previously, but is greater in children than for adults (3.38%) (Manjón and Coupé, 2016). *volBrain* segmentation uses manually labeled brain templates from 50 individuals with ages from 2 years old and 24–80 years old (Manjón and Coupé, 2016), which have no overlap with the age range of the current study (6–19 years old). The opposite directions of the estimation differences between *FreeSurfer* and *volBrain* methods imply, other than using unmatched templates, the vast variation in automatic extraction that may exist. Further studies are clearly warranted to explore whether the use of age-matched templates could improve the accuracy of automatic amygdala segmentation (Dong et al., 2020). Given the systematic differences in the amygdala volume between automatic and manual segmentation, it calls for caution on interpreting the results of the absolute amygdala volumes obtained by using the automatic methods in children and adolescents.

FreeSurfer and *FSL* exhibited more spatial volume overlap than *volBrain* with the manual tracing method. The spatial overlap (76–79%) observed between automatic methods (*CS-FS* and *FSL*) and the manual segmentation is consistent with the results reported by Morey et al. (2009). A higher overlap of *volBrain* was reported in a previous study (Manjón and Coupé, 2016), which is inconsistent with the observation in the present work. This could be related to the excessive underestimation of volume caused by the age-mismatched brain templates used by *volBrain* when segmenting amygdala for children and adolescents. In terms of spatial overlap, *FreeSurfer* and *FSL* outperformed *volBrain* for human pediatric amygdala segmentation. However, in terms of false-positive rate, *FreeSurfer* performed less than *FSL*, which in turn less than *volBrain*. The high false-positive rate of *FreeSurfer* could be an indication of its overestimation of the volume. A previous study suggested that it was due to excessive segmentation of brain structures in *FreeSurfer* by including structures and areas not part of the target structure (Næss-Schmidt et al., 2016). Correspondingly, the overestimation of *FreeSurfer* help to reduce errors of the exclusion of larger proportions of amygdalar structure, which resulting lower false-negative rates than *volBrain*. In terms of false-negative rate, *FreeSurfer* performed best among these automatic methods. Although few studies have explored the performance of *volBrain* on human amygdala segmentation in terms of the false-positive rates and false-negative rates, similar performance results have been shown for the automatic segmentation of the hippocampus and thalamic volume (Næss-Schmidt et al., 2016). According to inter-individual differences in segmented amygdala volumes, the two automatic methods only demonstrated moderate correlation with the manual segmentation, while *FSL* was not

significantly correlated with manual tracing. These are consistent with previous work (Morey et al., 2009; Grimm et al., 2015), implying its potential challenge for reliable measurements of their growth curves. Though, *LG-FS* showed slightly lower (not statistically significant) false-negative rates than the *CS-FS* segmentation, which outperformed *LG-FS* in terms of volume difference, spatial overlap and false positive rate metrics. The possible bias by matching head sizes across all the time points in children caused the worse accuracy of *LG-FS* than *CS-FS*. Overall, the *CS-FS*, *volBrain* and *FSL* methods have advantages and disadvantages for the assessment of amygdala volume. The complex amygdala structure adds difficulty to reliably and validly estimate its volume. It is a trade-off to choose which method should be used, requiring careful evaluation, and also demonstrates which facet of the automatic methods should be further improved in the future.

In this study, we found that the automatic segmentation performed worse in smaller amygdalae in developmental neuroimaging studies of children and adolescents. After controlling the quality of scan, the segmentation accuracy increased with amygdala volume, and then remained stable when the amygdala has reached a large enough size (around 1.3–1.4 ml). Previous studies have found that smaller brain structures were associated with greater automatic segmentation errors due to their sizes and shapes differing from adults (Schoemaker et al., 2016; Biffen et al., 2020; Sánchez-Benavides et al., 2010). Our results are consistent with that neuro-anatomical and geometric features could systematically influence the accuracy of their automatic segmentation. This bias is likely less problematic in adults, whose structures are commonly larger than in children. Poor scan quality caused less precise differentiation and classification of amygdala structures when using *FreeSurfer* and *volBrain* (see Supplementary TableS1). However, after controlling the quality of scans, the segmentation accuracy was still improved when the amygdala volume increased, and then remained stable when the amygdala reached a large enough size. The patterns of the spatial overlap function were highly similar between the two models (i.e., with/without controlling CJV and SNR). Therefore, we considered the nonlinearity presented in Fig. 3 is more likely attributed to the use of age-unmatched brain templates and the anatomical complexity of human amygdala. Previous studies on the amygdala have reported that the human amygdala could undergo the significant and complex deformation during childhood and adolescence (Schoemaker et al., 2016). The adult templates used in the automatic protocols mismatched with the current developing samples, especially with the young children, more likely causing poor segmentation results of young amygdalae. The human amygdala has been widely investigated in pediatric studies and associated with many developmental disorders such as autism (Mosconi et al., 2009; Schumann et al., 2009, 2004) and anxiety disorder (De Bellis et al., 2000; Hill et al., 2010; Milham et al., 2005). Our findings further highlighted the importance of using age-matched template and improving the measurement accuracy of automatic segmentation for developing individuals. We argue that, in the current stage, manual tracing should be given priority for amygdala volume estimation in pediatric research. In the future, eliminating the bias in automatic segmentation methods will be of great importance.

Although the statistical models indicated that the systematic differences in amygdala volume exhibited moderately marginal effects on growth curve modeling between the *FSL* and manual segmentation, our post-hoc growth chart analyses demonstrated remarkable discrepancies in age-related changes of the human amygdala across development. As the 'gold standard', manually traced amygdala volumes exhibited linear growth patterns without sex differences in growth rate. This is completely consistent with the patterns validated by the manual tracing method for amygdala growth from youth to adulthood in the macaque monkey (Schumann et al., 2019). Most previous studies of amygdala development in children and adolescents have been based on automatic segmentations (Wierenga et al., 2014; Goddings et al., 2014; Herting et al., 2018; Uematsu et al., 2012) while manual segmentation has been used in only two studies (Giedd et al., 1996; Merke et al., 2003). We noted that the developmental patterns of the amygdala have been inconsistent across these studies between automatic and manual methods. The growth patterns we detected by manual tracing are generally consistent with that by Giedd et al. (1996) and Merke et al. (2003) although they observed volume increases only in boys, but not in girls. In our study, the amygdala volumes grew in both boys and girls along highly similar trajectories. Such distinction may be an indication of the difference in scanning field strengths (3T versus 1.5T). Higher-resolution MRI enabled us to detect subtle changes in the human amygdala volume. Regarding automatic segmentation, previous studies generated amygdala growth curves with inverted U shapes from childhood to adolescence with peaks around 12–15 years old (Wierenga et al., 2014; Goddings et al., 2014; Herting et al., 2018; Uematsu et al., 2012). These were similar to our findings based on *CS-FS* and *FSL* segmentation, which showed a nonlinear trend of growth, especially for the right amygdala, with an inverted U-shaped trajectory (the volume peak at 14.18 years old). *LG-FS* produced similar shape with those of *CS-FS*, all exhibiting somehow nonlinearity although its degree of nonlinearity is left–right flipped between the two FS segmentation methods (Fig. 4). Although the *LG-FS* detected the statistical significance of amygdala's age-related increases, it achieved by overestimating amygdala volume more than *CS-FS* and sacrificing its segmentation accuracy. Therefore, only in the case of exploring the growth trend of the amygdala, *LG-FS* is preferable to *CS-FS*. *volBrain* segmentation yielded growth curves most similar to that obtained by the manual tracing for the amygdala development. *volBrain* seems to have less error modeling growth curves than *FreeSurfer*. However, given limited studies using *volBrain* to investigate amygdala development in children and adolescents, it is hard to compare our results with others directly.

The growth curves between the automatic methods (except for *FSL*) and manual tracing became similar when we adjusted amygdala volumes by the total gray matter volume rather than intracranial volume. This may reflect the reduction in the bias related to the amygdala size in automatic segmentation as mentioned above correcting the amygdala volume. As shown in Figure S7, the GMV derived with all methods (except for *FSL*) decreased as growing. The inclusion of GMV as a covariate in regression of amygdala volume on age could explain some variances of the amygdalar volume decreases (ei.ge., the decreasing part of the inverted-U shapes), resulting in the changes of developmental patterns from weak linearity to strong linearity as we demonstrated. We also found that a smaller GMV was associated with worse performance of automatic amygdala segmentation (except for *FSL*) but remain stable for large enough GMVs (Figure S4). As the GMV decreased with age, the measurement bias for younger children could be removed more than older children by controlling for GMV. This further strengthened the linearity of the developmental patterns of *volBrain* and *FreeSurfer*. In contrast, the measurement bias of *FSL* was negatively related to GMV, leading to more corrections of the measurement bias for older participants and thus the aggravated decreases. As shown in Figure S8, the ICV obtained with the three methods all increased when growing. The inclusion of ICV as a covariate in regression could explain some variances of age-related increases of amygdala volume. *volBrain* demonstrated a segmentation bias associated with ICV (Figure S5), indicating a smaller brain associated with worse amygdala segmentation but remain stable for big enough brains. This might cause the changes of developmental patterns of *volBrain*-derived amygdala from strong linearity to weak linearity. However, *CS-FS* did not demonstrate such an ICV-related segmentation bias (Figure S5) but higher false positive rates of segmentation associated with larger brains (Figure S6). After controlling ICV, the amygdala volumes of the smaller brains were corrected little, while the volume estimates of the larger brains were corrected more. This might explain that the changing pattern of *CS-FS* derived growth curves from increasing with age to decreasing with age. These results suggest that controlling for the gray matter volume

improved the accuracy of curve-fitting on the *volBrain* and *FreeSurfer* of amygdala from childhood to adolescence.

Accurate delineation of the development of the human amygdala is fundamentally important by providing neuroimaging biomarkers for various developmental disorders (DiMartino et al., 2014; Zuo, 2020; Holla et al., 2020). Our findings present an unaddressed bias and challenge for charting the growth of the human amygdala across school-age children and adolescents-the growth curve modeling was highly dependent on the segmentation method. The methodological differences may contribute to the inconsistencies among previous findings regarding the patterns of amygdala development during childhood and adolescence (Wierenga et al., 2018; Uematsu et al., 2012; Albaugh et al., 2017; Herting et al., 2018). Given the inconsistency, we give researchers working on the amygdala of children and adolescents some suggestions: (1) Manually tracing the amygdala if possible. This seems affordable but prohibitive for very large-scale MRI datasets, considering it takes 90 min to manually tracing an amygdala and will take 3 months to manually trace 500 amygdalae. A practical solution would be distributing the mission to a tracing team achieved high within-operator and between-operator reliability of the tracing operation by an established training protocol. (2) Checking and correcting the automatic segmentation of the amygdala by a trained professional to improve the accuracy and save the effort if the manual segmentation is not feasible. This would likely lead to significantly improved accuracy and time cost of the tracing segmentation although need further investigations in future. In addition, a very promising solution is to train a computational segmentation tool by integrating the knowledge aggregated from big data of the manually segmented amygdalae by using some novel methods (e.g., the deep learning algorithms). (3) Using age-matched brain templates for automatic segmentation (Dong et al., 2020). (4) Using high-resolution MRI protocol on scanning the amygdala. (5) Adjusting the amygdala volume by total gray matter volume when conducting statistical analysis. (6) Comparing and interpreting previous findings cautiously using different segmentation methods than the study proposed, particularly for the smaller amygdalae (e.g., younger children and females). To facilitate the use of the growth curves for human amygdala development at school age, we made the manually traced amygdalae publicly available to the community via the National Science Data Bank (http://www.doi.org/10.11922/sciencedb.01299).

Our study has some limitations that should be noted. First, the age span of our sample might not be sufficient for examining the full range of development of the human amygdala from childhood, adolescence and into young adulthood. While the previous work in the macaque monkey revealed the linear pattern of amygdala growth from youth to adulthood (Schumann et al., 2019), further work would benefit from the extension of the age span into adulthood for direct growth assessments in human in future. Second, we did not investigate the measurement reliability across different versions of automatic segmentation tools, which has been shown remarkable influences on the brain segmentation (Gronenschild et al., 2012). This factor should be carefully evaluated by using different versions of these tools to model amygdala growth. Finally, we only examined the overall volume measurement of the human amygdala. In the future, we will employ more local and shape measurements (Li et al., 2012; Roshchupkin et al., 2016) for investigating more details of human amygdala growth. To provide more efficient and accurate tracing of the pediatric amygdala, we also plan to develop an automatic algorithm based upon the manually traced samples using more advanced methods such as deep learning (Ataloglou et al., 2019).

5. Conclusion

By manually tracing a large-sample pediatric MRI dataset from the accelerated longitudinal cohort, we charted the growth of human amygdala across school age. We identified measurement biases for the automatic amygdala segmentation methods and their impacts on modeling growth curves of the amygdala volumes from childhood to adolescence. There is considerable room for the methodological improvement on next generation tools of automatic segmentation to achieve more accurately tracing of the human amygdala during development. Our work provides not only a practical guideline for future studies on amygdala in children and adolescents but also its growth standard resources for translational and educational applications. This can be implemented with normative modeling (Holla et al., 2020; Marquand et al., 2016; Liu et al., 2021) for individualized assessments on typical or atypical development as well as their associations with behavioral performance, school achievement and clinical symptoms.

6. The chinese color nest consortium

The Chinese Color Nest Consortium members are at http://deepneuro.bnu.edu.cn/?p=163.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data statement

A set of the baseline devCCNP-SWU data on brain imaging has been made available to researchers via the CoRR, which is committed to open science by aggregating and sharing MRI data from multiple sources to establish test-retest reliability and reproducibility in functional connectomics (http://dx.doi.org/10.15387/fcp_indi.corr.ipcas7). The rest of the data will be publicly shared via the National Science Data Bank and fully available to the research community when acquisition is completed for the pilot CCNP cohort. At this stage, data are only available to researchers and collaborators of CCNP. More information about CCNP can be found at: http://deepneuro.bnu.edu.cn/?p=163 or https://github.com/zuoxinian/CCNP. Requests for further information and collaboration are encouraged and considered by principal investigator Xi-Nian Zuo [xinian.zuo@bnu.edu.cn].

Acknowledgments

We thank all families participating in the Chinese Color Nest Program and all the support from schools and communities. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province, China (2019B030335001), the Start-up Funds for Leading Talents at Beijing Normal University, the National Basic Science Data Center "Chinese Data-sharing Warehouse for In-vivo Imaging Brain" (NBSDC-DB-15), the Beijing Municipal Science and Technology Commission, China (Z161100002616023, Z181100001518003), the CAS-NWO Programme (153111KYSB20160020), the National Natural Science Foundation of China (81220108014), the Guangxi BaGui Scholarship, China (201621) and National Basic Research (973) Program, China (2015CB351702). Finally, we would like to thank Tonya White, MD, PhD and Neda Jahanshad, PhD for their comments, suggestions, and for proofing the article for English language.

Appendix A. Supplementary figures and tables

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.dcn.2021.101028.

References

Akudjedu, T.N., Nabulsi, L., Makelyte, M., Scanlon, C., Hehir, S., Casey, H., Ambati, S., Kenney, J., O'Donoghue, S., McDermott, E., et al., 2018. A comparative study of segmentation techniques for the quantification of brain subcortical volume. Brain Imaging Behav. 12 (6), 1678–1695. http://dx.doi.org/10.1007/s11682-018-9835-y.

Albaugh, M.D., Nguyen, T.-V., Ducharme, S., Collins, D.L., Botteron, K.N., D'Alberto, N., Evans, A.C., Karama, S., Hudziak, J.J., Group, B.D.C., et al., 2017. Age-related volumetric change of limbic structures and subclinical anxious/depressed symptomatology in typically developing children and adolescents. Biol. Psychol. 124, 133–140. http://dx.doi.org/10.1016/j.biopsycho.2017.02.002.

Alexander-Bloch, A.F., Reiss, P.T., Rapoport, J., McAdams, H., Giedd, J.N., Bullmore, E.T., Gogtay, N., 2014. Abnormal cortical growth in schizophrenia targets normative modules of synchronized development. Biol. Psychiat. 76 (6), 438–446. http://dx.doi.org/10.1016/j.biopsych.2014.02.010.

Ataloglou, D., Dimou, A., Zarpalas, D., Daras, P., 2019. Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning. Neuroinformatics 17 (4), 563–582. http://dx.doi.org/10.1007/s12021-019-09417-y.

Barnea-Goraly, N., Frazier, T.W., Piacenza, L., Minshew, N.J., Keshavan, M.S., Reiss, A.L., Hardan, A.Y., 2014. A preliminary longitudinal volumetric MRI study of amygdala and hippocampal volumes in autism. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 48, 124–128. http://dx.doi.org/10.1016/j.pnpbp.2013.09.010.

Biffen, S.C., Warton, C.M., Dodge, N.C., Molteno, C.D., Jacobson, J.L., Jacobson, S.W., Meintjes, E.M., 2020. 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. Neuroimage: Clin. 28, 102368. http://dx.doi.org/10.1016/j.nicl.2020.102368.

Brain Development Cooperative Group, 2012. Total and regional brain volumes in a population-based normative sample from 4 to 18 years: the NIH mri study of normal brain development. Cereb. Cortex 22 (1), 1–12. http://dx.doi.org/10.1093/cercor/bhr018.

Brown, C.J., Miller, S.P., Booth, B.G., Andrews, S., Chau, V., Poskitt, K.J., Hamarneh, G., 2014. Structural network analysis of brain development in young preterm neonates. Neuroimage 101, 667–680. http://dx.doi.org/10.1016/j.neuroimage.2014.07.030.

Cardinal, R.N., Parkinson, J.A., Hall, J., Everitt, B.J., 2002. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci. Biobehav. Rev. 26 (3), 321–352. http://dx.doi.org/10.1016/s0149-7634(02)00007-6.

Connor, D.F., 2004. Aggression and Antisocial Behavior in Children and Adolescents: Research and Treatment. Guilford Press.

De Bellis, M.D., Casey, B., Dahl, R.E., Birmaher, B., Williamson, D.E., Thomas, K.M., Axelson, D.A., Frustaci, K., Boring, A.M., Hall, J., et al., 2000. A pilot study of amygdala volumes in pediatric generalized anxiety disorder. Biol. Psychiat. 48 (1), 51–57. http://dx.doi.org/10.1016/s0006-3223(00)00835-0.

DiMartino, A., Fair, D., Kelly, C., Satterthwaite, T., Castellanos, F., Thomason, M., Craddock, R., Luna, B., Leventhal, B., Zuo, X.-N., Milham, M., 2014. Unraveling the miswired connectome: A developmental perspective. Neuron 83 (6), 1335–1353. http://dx.doi.org/10.1016/j.neuron.2014.08.050.

Dong, H.-M., Castellanos, F.X., Yang, N., Zhang, Z., Zhou, Q., He, Y., Zhang, L., Xu, T., Holmes, A., Yeo, B.T.T., Chen, F., Wang, B., Beckmann, C., White, T., Sporns, O., Qiu, J., Feng, T., Chen, A., Liu, X., Chen, X., Weng, X., Milham, M.P., Zuo, X.-N., 2020. Charting brain growth in tandem with brain templates for schoolchildren. Sci. Bull. 65, 1924–1934. http://dx.doi.org/10.1016/j.scib.2020.07.027.

Dong, H.-M., Margulies, D.S., Zuo, X.-N., Holmes, A.J., 2021. Shifting gradients of macroscale cortical organization mark the transition from childhood to adolescence. Proc. Natl. Acad. Sci. USA 118, e2024448118. http://dx.doi.org/10.1073/pnas.2024448118.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355. http://dx.doi.org/10.1016/s0896-6273(02)00569-x.

Ganzetti, M., Wenderoth, N., Mantini, D., 2016. Intensity inhomogeneity correction of structural MR images: a data-driven approach to define input algorithm parameters. Front. Neuroinform. 10, 10. http://dx.doi.org/10.3389/fninf.2016.00010.

Ganzola, R., Maziade, M., Duchesne, S., 2014. Hippocampus and amygdala volumes in children and young adults at high-risk of schizophrenia: research synthesis. Schizophr. Res. 156 (1), 76–86. http://dx.doi.org/10.1016/j.schres.2014.03.030.

Giedd, J.N., Vaituzis, A.C., Hamburger, S.D., Lange, N., Rajapakse, J.C., Kaysen, D., Vauss, Y.C., Rapoport, J.L., 1996. Quantitative MRI of the temporal lobe, amygdala, and hippocampus in normal human development: ages 4–18 years. J. Comp. Neurol. 366 (2), 223–230. http://dx.doi.org/10.1002/(SICI)1096-9861(19960304)366:2<223::AID-CNE3>3.0.CO;2-7.

Gilmore, J.H., Shi, F., Woolson, S.L., Knickmeyer, R.C., Short, S.J., Lin, W., Zhu, H., Hamer, R.M., Styner, M., Shen, D., 2012. Longitudinal development of cortical and subcortical gray matter from birth to 2 years. Cereb. Cortex 22 (11), 2478–2485. http://dx.doi.org/10.1093/cercor/bhr327.

Goddings, A.-L., Mills, K.L., Clasen, L.S., Giedd, J.N., Viner, R.M., Blakemore, S.-J., 2014. The influence of puberty on subcortical brain development. Neuroimage 88, 242–251. http://dx.doi.org/10.1016/j.neuroimage.2013.09.073.

Grimm, O., Pohlack, S., Cacciaglia, R., Winkelmann, T., Plichta, M.M., Demirakca, T., Flor, H., 2015. Amygdalar and hippocampal volume: a comparison between manual segmentation, freesurfer and VBM. J. Neurosci. Methods 253, 254–261. http://dx.doi.org/10.1016/j.jneumeth.2015.05.024.

Gronenschild, E., Habets, P., Jacobs, H., 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 (6), e38234. http://dx.doi.org/10.1371/journal.pone.0038234.

Harezlak, J., Ryan, L.M., Giedd, J.N., Lange, N., 2005. Individual and population penalized regression splines for accelerated longitudinal designs. Biometrics 61 (4), 1037–1048. http://dx.doi.org/10.1111/j.1541-0420.2005.00376.x.

He, Y., Xu, T., Zhang, W., Zuo, X.-N., 2016. Lifespan anxiety is reflected in human amygdala cortical connectivity. Hum. Brain Mapp. 37 (3), 1178–1193. http://dx.doi.org/10.1002/hbm.23094.

Herten, A., Konrad, K., Krinzinger, H., Seitz, J., von Polier, G.G., 2019. Accuracy and bias of automatic hippocampal segmentation in children and adolescents. Brain Struct. Funct. 224 (2), 795–810. http://dx.doi.org/10.1007/s00429-018-1802-2.

Herting, M.M., Gautam, P., Spielberg, J.M., Kan, E., Dahl, R.E., Sowell, E.R., 2014. The role of testosterone and estradiol in brain volume changes across adolescence: a longitudinal structural MRI study. Hum. Brain Mapp. 35 (11), 5633–5645. http://dx.doi.org/10.1002/hbm.22575.

Herting, M.M., Johnson, C., Mills, K.L., Vijayakumar, N., Dennison, M., Liu, C., Goddings, A.-L., Dahl, R.E., Sowell, E.R., Whittle, S., et al., 2018. Development of subcortical volumes across adolescence in males and females: A multisample study of longitudinal changes. Neuroimage 172, 194–205. http://dx.doi.org/10.1016/j.neuroimage.2018.01.020.

Hill, S.Y., Tessner, K., Wang, S., Carter, H., McDermott, M., 2010. Temperament at 5 years of age predicts amygdala and orbitofrontal volume in the right hemisphere in adolescence. Psychiatry Res. 182 (1), 14–21. http://dx.doi.org/10.1016/j.pscychresns.2009.11.006.

Holla, B., Seidlitz, J., Bethlehem, R., Schumann, G., 2020. Population normative models of human brain growth across development. Sci. Bull. 65 (22), 1872–1873. http://dx.doi.org/10.1016/j.scib.2020.08.040.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

LeDoux, J., 1998. The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon and Schuster.

Li, S., Wang, Y., Xu, P., Pu, F., Li, D., Fan, Y., Gong, G., Luo, Y., 2012. Surface morphology of amygdala is associated with trait anxiety. PLoS One 7 (10), e47817. http://dx.doi.org/10.1371/journal.pone.0047817.

Liu, S., Wang, Y.-S., Zhang, Q., Zhou, Q., Cao, L.-Z., Jiang, C., Zhang, Z., Yang, N., Dong, Q., Zuo, X.-N., Chinese Color Nest Consortium, T., 2021. Chinese color nest project: an accelerated longitudinal brain-mind cohort. Dev. Cogn. Neurosci. 52, 101020. http://dx.doi.org/10.1016/j.dcn.2021.101020.

Lyden, H., Gimbel, S.I., Del Piero, L., Tsai, A.B., Sachs, M.E., Kaplan, J.T., Margolin, G., Saxbe, D., 2016. Associations between family adversity and brain volume in adolescence: Manual vs. automated brain segmentation yields different results. Front. Neurosci. 10, 398. http://dx.doi.org/10.3389/fnins.2016.00398.

Manjón, J.V., Coupé, P., 2016. Volbrain: an online MRI brain volumetry system. Front. Neuroinform. 10, 30. http://dx.doi.org/10.3389/fninf.2016.00030.

Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F., 2016. Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. Biol. Psychiat. 80 (7), 552–561. http://dx.doi.org/10.1016/j.biopsych.2015.12.023.

Merke, D.P., Fields, J.D., Keil, M.F., Vaituzis, A.C., Chrousos, G.P., Giedd, J.N., 2003. Children with classic congenital adrenal hyperplasia have decreased amygdala volume: potential prenatal and postnatal hormonal effects. J. Clin. Endocrinol. Metab. 88 (4), 1760–1765. http://dx.doi.org/10.1210/jc.2002-021730.

Milchenko, M., Marcus, D., 2013. Obscuring surface anatomy in volumetric imaging data. Neuroinformatics 11 (1), 65–75. http://dx.doi.org/10.1007/s12021-012-9160-3.

Milham, M.P., Nugent, A.C., Drevets, W.C., Dickstein, D.S., Leibenluft, E., Ernst, M., Charney, D., Pine, D.S., 2005. Selective reduction in amygdala volume in pediatric anxiety disorders: a voxel-based morphometry investigation. Biol. Psychiat. 57 (9), 961–966. http://dx.doi.org/10.1016/j.biopsych.2005.01.038.

Mills, K.L., Tamnes, C.K., 2014. Methods and considerations for longitudinal structural brain imaging analysis across development. Dev. Cogn. Neurosci. 9, 172–190. http://dx.doi.org/10.1016/j.dcn.2014.04.004.

Morey, R.A., Petty, C.M., Xu, Y., Hayes, J.P., Wagner II, 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 (3), 855–866. http://dx.doi.org/10.1016/j.neuroimage.2008.12.033.

Mosconi, M.W., Cody-Hazlett, H., Poe, M.D., Gerig, G., Gimpel-Smith, R., Piven, J., 2009. Longitudinal study of amygdala volume and joint attention in 2-to 4-year-old children with autism. Arch. Gen. Psychiatry 66 (5), 509–516. http://dx.doi.org/10.1001/archgenpsychiatry.2009.19.

Næss-Schmidt, E., Tietze, A., Blicher, J.U., Petersen, M., Mikkelsen, I.K., Coupé, P., Manjón, J.V., Eskildsen, S.F., 2016. Automatic thalamus and hippocampus segmentation from MP2rage: comparison of publicly available methods and implications for dti quantification. Int. J. Comput. Assisted Radiol Surg. 11 (11), 1979–1991. http://dx.doi.org/10.1007/s11548-016-1433-0.

Narvacan, K., Treit, S., Camicioli, R., Martin, W., Beaulieu, C., 2017. Evolution of deep gray matter volume across the human lifespan. Hum. Brain Mapp. 38 (8), 3771–3790. http://dx.doi.org/10.1002/hbm.23604.

Nooner, K., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., Panek, L., Brown, S., Zavitz, S., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R., Kamiel, S., Anwar, A., Hinz, C., Kaplan, M., Rachlin, A., Adelsberg, S., Cheung, B., Khanuja, R., Yan, C., Craddock, C., Calhoun, V., Courtney, W., King, M., Wood, D., Cox, C., Kelly, A., Di Martino, A., Petkova, E., Reiss, P., Duan, N., Thomsen, D., Biswal, B., Coffey, B., Hoptman, M., Javitt, D., Pomara, N., Sidtis, J., Koplewicz, H., Castellanos, F., Leventhal, B., Milham, M., 2012. The NKI-rockland sample: A model for accelerating the pace of discovery science in psychiatry. Front. Neurosci. 6, 152. http://dx.doi.org/10.3389/fnins.2012.00152.

Ortiz, J., Raine, A., 2004. Heart rate level and antisocial behavior in children and adolescents: A meta-analysis. J. Amer. Acad. Child Adolesc. Psychiatry 43 (2), 154–162. http://dx.doi.org/10.1097/00004583-200402000-00010.

Patenaude, B., Smith, S., Kennedy, D., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56 (3), 907–922. http://dx.doi.org/10.1016/j.neuroimage.2011.02.046.

Paus, T., Keshavan, M., Giedd, J., 2008. Why do many psychiatric disorders emerge during adolescence? Nat. Rev. Neurosci. 9 (12), 947–957. http://dx.doi.org/10.1038/nrn2513.

Pessoa, L., 2010. Emotion and cognition and the amygdala: from "what is it?" to "what's to be done?". Neuropsychologia 48 (12), 3416–3429. http://dx.doi.org/10.1016/j.neuropsychologia.2010.06.038.

Pruessner, J.C., Li, L.M., Serles, W., Pruessner, M., Collins, D.L., Kabani, N., Lupien, S., Evans, A.C., 2000. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb. Cortex 10 (4), 433–442. http://dx.doi.org/10.1093/cercor/10.4.433.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

Redlich, R., Grotegerd, D., Opel, N., Kaufmann, C., Zwitserlood, P., Kugel, H., Heindel, W., Donges, U.-S., Suslow, T., Arolt, V., et al., 2015. Are you gonna leave me? Separation anxiety is associated with increased amygdala responsiveness and volume. Soc. Cogn. Affect. Neurosci. 10 (2), 278–284. http://dx.doi.org/10.1093/scan/nsu055.

Reuter, M., Schmansky, N., Rosas, H., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418. http://dx.doi.org/10.1016/j.neuroimage.2012.02.084.

Rice, K., Viscomi, B., Riggins, T., Redcay, E., 2014. Amygdala volume linked to individual differences in mental state inference in early childhood and adulthood. Dev. Cogn. Neurosci. 8, 153–163. http://dx.doi.org/10.1016/j.dcn.2013.09.003.

Roshchupkin, G., Gutman, B., Vernooij, M., Jahanshad, N., Martin, N., Hofman, A., McMahon, K., Van Der Lee, S., Van Duijn, C., De Zubicaray, G., Uitterlinden, A., Wright, M., Niessen, W., Thompson, P., Ikram, M., Adams, H., 2016. Heritability of the shape of subcortical brain structures in the general population. Nat. Comms. 7, 13738. http://dx.doi.org/10.1038/ncomms13738.

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 (3), 219–225. http://dx.doi.org/10.1016/j.pscychresns.2009.10.011.

Sawiak, S., Shiba, Y., Oikonomidis, L., Windle, C., Santangelo, A.M., Grydeland, H., Cockcroft, G., Bullmore, E., Roberts, A., 2018. Trajectories and milestones of cortical and subcortical development of the marmoset brain from infancy to adulthood. Cereb. Cortex 28 (12), 4440–4453. http://dx.doi.org/10.1093/cercor/bhy256.

Scherf, K.S., Smyth, J.M., Delgado, M.R., 2013. The amygdala: an agent of change in adolescent neural networks. Horm. Behav. 64 (2), 298–313. http://dx.doi.org/10.1016/j.yhbeh.2013.05.011.

Schmidt, M.F., Storrs, J.M., Freeman, K.B., Jack Jr., C.R., Turner, S.T., Griswold, M.E., Mosley Jr., T.H., 2018. A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan. Hum. Brain Mapp. 39 (6), 2500–2513. http://dx.doi.org/10.1002/hbm.24017.

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. http://dx.doi.org/10.1016/j.neuroimage.2016.01.038.

Schumann, C.M., Amaral, D.G., 2005. Stereological estimation of the number of neurons in the human amygdaloid complex. J. Comp. Neurol. 491 (4), 320–329. http://dx.doi.org/10.1002/cne.20704.

Schumann, C.M., Barnes, C.C., Lord, C., Courchesne, E., 2009. Amygdala enlargement in toddlers with autism related to severity of social and communication impairments. Biol. Psychiat. 66 (10), 942–949. http://dx.doi.org/10.1016/j.biopsych.2009.07.007.

Schumann, C.M., Bauman, M.D., Amaral, D.G., 2011. Abnormal structure or function of the amygdala is a common component of neurodevelopmental disorders. Neuropsychologia 49 (4), 745–759. http://dx.doi.org/10.1016/j.neuropsychologia.2010.09.028.

Schumann, C.M., Hamstra, J., Goodlin-Jones, B.L., Lotspeich, L.J., Kwon, H., Buonocore, M.H., Lammers, C.R., Reiss, A.L., Amaral, D.G., 2004. The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages. J. Neurosci. 24 (28), 6392–6401. http://dx.doi.org/10.1523/JNEUROSCI.1297-04.2004.

Schumann, C., Scott, J., Lee, A., Bauman, M., Amaral, D., 2019. Amygdala growth from youth to adulthood in the macaque monkey. J. Comp. Neurol. 527 (18), 3034–3045. http://dx.doi.org/10.1002/cne.24728.

Silk, J.S., Vanderbilt-Adriance, E., Shaw, D.S., Forbes, E.E., Whalen, D.J., Ryan, N.D., Dahl, R.E., 2007. Resilience among children and adolescents at risk for depression: Mediation and moderation across social and neurobiological contexts. Dev. Psychopathol. 19 (3), 841–865. http://dx.doi.org/10.1017/S0954579407000417.

Tamnes, C.K., Walhovd, K.B., Dale, A.M., Østby, Y., Grydeland, H., Richardson, G., Westlye, L.T., Roddey, J.C., Hagler Jr., D.J., Due-Tønnessen, P., et al., 2013. Brain development and aging: overlapping and unique patterns of change. Neuroimage 68, 63–74. http://dx.doi.org/10.1016/j.neuroimage.2012.11.039.

Thompson, W., Hallmayer, J., O'Hara, R., 2011. Design considerations for characterizing psychiatric trajectories across the life span: Application to effects of APOE-ϵ4 on cerebral cortical thickness in alzheimer's disease. Am. J. Psychiatry 168 (9), 894–903. http://dx.doi.org/10.1176/appi.ajp.2011.10111690.

Uematsu, A., Matsui, M., Tanaka, C., Takahashi, T., Noguchi, K., Suzuki, M., Nishijo, H., 2012. Developmental trajectories of amygdala and hippocampus from infancy to early adulthood in healthy individuals. PLoS One 7 (10), e46970. http://dx.doi.org/10.1371/journal.pone.0046970.

Van Petten, C., 2004. Relationship between hippocampal volume and memory ability in healthy individuals across the lifespan: review and meta-analysis. Neuropsychologia 42 (10), 1394–1413. http://dx.doi.org/10.1016/j.neuropsychologia.2004.04.006.

Watson, C., Andermann, F., Gloor, P., Jones-Gotman, M., Peters, T., Evans, A., Olivier, A., Melanson, D., Leroux, G., 1992. Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. Neurology 42 (9), 1743. http://dx.doi.org/10.1212/wnl.42.9.1743.

Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 8 (11), e77089. http://dx.doi.org/10.1371/journal.pone.0077089.

Wickham, H., 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, URL https://ggplot2.tidyverse.org.

Wierenga, L.M., Bos, M.G., Schreuders, E., vd Kamp, F., Peper, J.S., Tamnes, C.K., Crone, E.A., 2018. Unraveling age, puberty and testosterone effects on subcortical brain development across adolescence. Psychoneuroendocrinology 91, 105–114. http://dx.doi.org/10.1016/j.psyneuen.2018.02.034.

Wierenga, L., Langen, M., Ambrosino, S., van Dijk, S., Oranje, B., Durston, S., 2014. Typical development of basal ganglia, hippocampus, amygdala and cerebellum from age 7 to 24. Neuroimage 96, 67–72. http://dx.doi.org/10.1016/j.neuroimage.2014.03.072.

Wood, S., 2017. Generalized Additive Models: An Introduction with R, second ed. Chapman and Hall/CRC.

Xing, X.-X., Zuo, X.-N., 2018. The anatomy of reliability: a must read for future human brain mapping. Sci. Bull. 63 (24), 1606–1607. http://dx.doi.org/10.1016/j.scib.2018.12.010.

Xu, T., Yang, Z., Jiang, L., Xing, X.-X., Zuo, X.-N., 2015. A connectome computation system for discovery science of brain. Sci. Bull. 60 (1), 86–95. http://dx.doi.org/10.1007/s11434-014-0698-3.

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

Zuo, X.-N., 2020. Editorial: Mapping the miswired connectome in autism spectrum disorder. J. Amer. Acad. Child Adolesc. Psychiatry 59 (3), 348–349. http://dx.doi.org/10.1016/j.jaac.2020.01.001.

Zuo, X.-N., He, Y., Betzel, R.F., Colcombe, S., Sporns, O., Milham, M.P., 2017. Human connectomics across the life span. Trends Cogn. Sci. 21 (1), 32–45. http://dx.doi.org/10.1016/j.tics.2016.10.005.