



Semi-automatic segmentation of the fetal brain from magnetic resonance imaging

Jianan Wang, Emily S. Nichols, Megan E. Mueller, Barbra de Vrijer, Roy Eagleson, Charles A. McKenzie, Sandrine de Ribaupierre and Emma G. Duerden

Frontier, November 2022



Background: Improved understanding of fetal brain development trajectories may aid in identifying and clinically managing at-risk fetuses.

Problem: fetal brain structures in magnetic resonance images (MRI) are often manually segmented, which requires both time and expertise

goal: To facilitate the targeting and measurement of brain structures in the fetus, we compared the results of five segmentation methods applied to fetal brain MRI data to gold-standard manual tracings.

Method: dataset: Adult women with singleton pregnancies ($n = 21$), of whom five were scanned twice, approximately 3 weeks apart, were recruited [26 total datasets]. T2-weighted single-shot fast spin echo images of the fetal brain were acquired on 1.5T and 3T MRI scanners. Images were first combined into a single 3D anatomical volume. Next, a trained tracer manually segmented the thalamus, cerebellum, and total cerebral volumes. The manual segmentations were compared with five automatic methods of segmentation available within Advanced Normalization Tools (ANTs) and FMRIB's Linear Image Registration Tool (FLIRT) toolboxes. The manual and automatic labels were compared using Dice similarity coefficients (DSCs). The DSC values were compared using Friedman's test for repeated measures.



Introduction

For functional imaging, manual segmentation of 4D fetal images can take upwards of 30 h to complete a single scanning run in an individual participant's data

problem: Automatic segmentation pipelines and routines developed for neonatal and child imaging protocols are not appropriate for studying fetal brain tissue due to the variations in image acquisition, and maturational differences leading to poorer contrast of the gray and white matter.



Introduction

segmentation techniques: They applied two atlas-based segmentation techniques, linear and nonlinear atlas registration algorithms, to perform the regional segmentation of the cortex and subcortical areas in the fetal brain to examine their macrostructural development.

Using an atlas-based method, they examined whether more computationally intensive deformation image registration methods, using the Advanced Normalization Tools (ANTs), are needed for adequate subcortical segmentation compared to an affine image registration FLIRT (FMRIB's Linear Image Registration Tool). This research aimed to develop and implement a semi-automatic pipeline combining semi-automatic fetal brain reconstruction, segmentation, volumetric reconstruction, and atlas registration algorithms for subcortical segmentation in fetal brains to extract and analyze subcortical volumes



Materials and methods



Volumetric reconstruction of magnetic resonance images

NiftyMIC (Ebner et al., 2020) was used for fetal brain segmentation and 3D reconstruction. The main processing pipeline for detection and segmentation of the fetal brain included with NiftyMIC involves only a single command (`fetal_brain_seg`) and can be executed unsupervised. Various features of different slice-to-volume reconstructions methods including NiftyMIC have been compared for fetal MRI, and have reported comparable results (Payette et al., 2021).

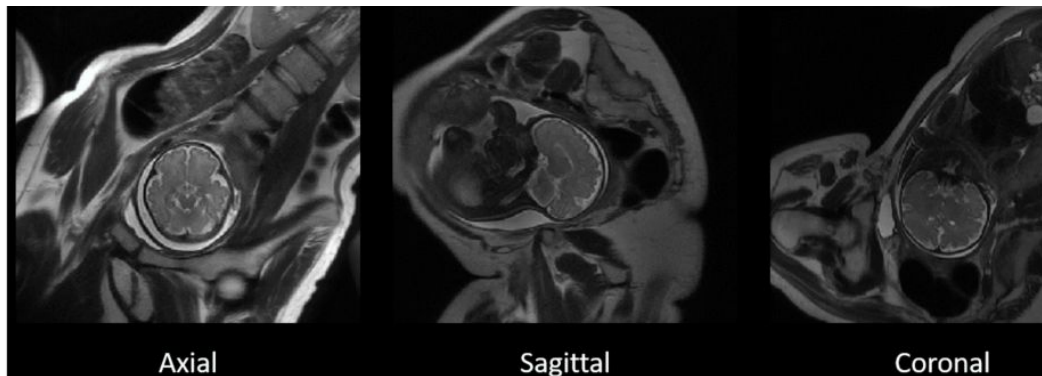
It was essential to estimate the fetal brain location in the MR image such that a bounding box was created to reduce both unrelated contents and image space, as well as the algorithm processing time for the later more precise fetal brain segmentation algorithm using 2D P-Net CNN.

NiftyMIC's `fetal_brain_seg` command was then executed on the MR image, generating a mask of the fetal brain in the surrounding tissue for each slice within the image. This step took under 2 min per stack of 2D slices.

The resulting masks were then reviewed using FSLeyes. These automatically generated 2D fetal brain masks from NiftyMIC were suboptimal for most participants, resulting in either over- or under-estimating fetal brain tissue in the slices; surrounding maternal gray and white tissue were still evident in the slices, depending on the acquisition and field of view. Therefore, manual adjustments of the masks, such as filling and excluding pixels, were performed on all automatically generated 2D masks ($n = 26$). Time spent manually editing ranged from 1 to 15 min per stack of 2D slices, with the majority taking under 5 min to complete.

Volumetric reconstruction of magnetic resonance images

The original T2-weighted acquisition of a fetal MR image in axial, sagittal, and coronal planes. T2-weighted images acquired separately in three separate image planes in the axial (left), sagittal (middle), and coronal (right) in a representative participant. The three image planes were subsequently used for the reconstruction of 3D images





Volumetric reconstruction of magnetic resonance images

After segmenting fetal brains in the 2D planes, the stacks of 2D slices were reconstructed into 3D volumes, and the 2D fetal brain segmentations were also reconstructed into 3D space.

1-The 2D MR image slices could be corrupted by low-frequency bias field signals to blur the high-frequency contents, such as edges and contours. Intensity variance also resulted from existing bias field signals where the same tissue had a uniformed pixel gray level in the images. ==> stacks of segmented 2D fetal brain slices were first bias-field corrected
2-slice-to-volume process: the bias-field corrected 2D slices were reconstructed into a 3D volume by the slice-to-volume process that rigidly registered the 2D slices to one randomly selected target slice from the fetal brain MR images so that all the slices were volumetrically aligned. The slice-to-volume process also used linear regression to correct and match the slices' voxel intensities to the target slice's voxel intensity.

3-volume-to-volume process: It was performed on the 2D slices and previously segmented 2D masks to reconstruct into 3D volumes and 3D fetal brain masks in native space. Processing times varied but were up to 2 h in some participants' data

4-Subsequently, the native-space 3D volumes were rigidly registered to a spatiotemporal atlas developed from images acquired at 3T MRI from typically developing fetuses to obtain the volumetric reconstruction in the standard anatomical planes of atlas space



Registration-based subcortical segmentation

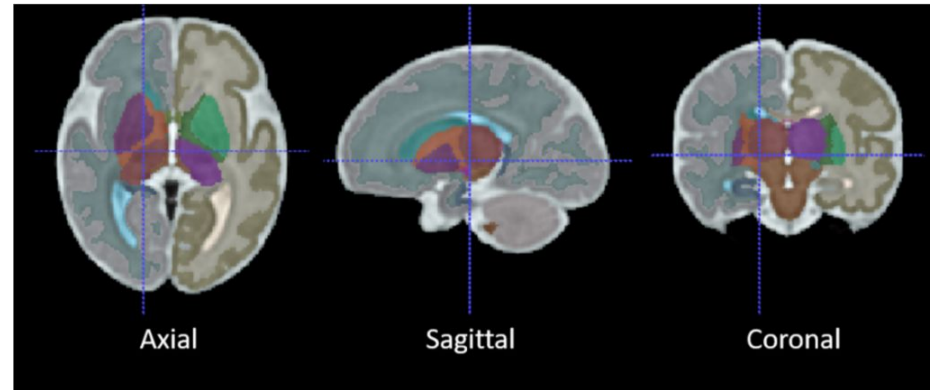
Skull stripping: The reconstructed 3D fetal brain masks were applied onto the reconstructed 3D brain volumes for fetal brain skull stripping. The 3D brain volumes were segmented with the binary masks for fetal brain-only MR images. (using 3dcalc from the AFNI toolkit that multiplied the reconstructed 3D fetal brain image with the binary 3D masks) -> Then the orientations of the skull-stripped MR images were manually adjusted according to the orientations of the age-appropriate fetal brain atlas using the ITK-SNAP GUI.

Two different registration toolkits were applied to reconstructed images and compared: 1- Deformable registration using ANTs using the well-known SyN (symmetric image normalization) method 2-linear (affine) atlas registration using FLIRT

The fetal brain atlas: is an averaged atlas from fetuses with predefined labels of deep-brain structures, including the thalamus and cerebellum. The atlas was nonlinearly and linearly registered into the native participant 3D MRI space. The transformation matrix was saved and applied onto the atlas mask to warp the tissue labels into subject space. The transformed atlas labels were used as thalamus and cerebellum masks and were compared with manual masks by calculating DSCs for the reliability test.

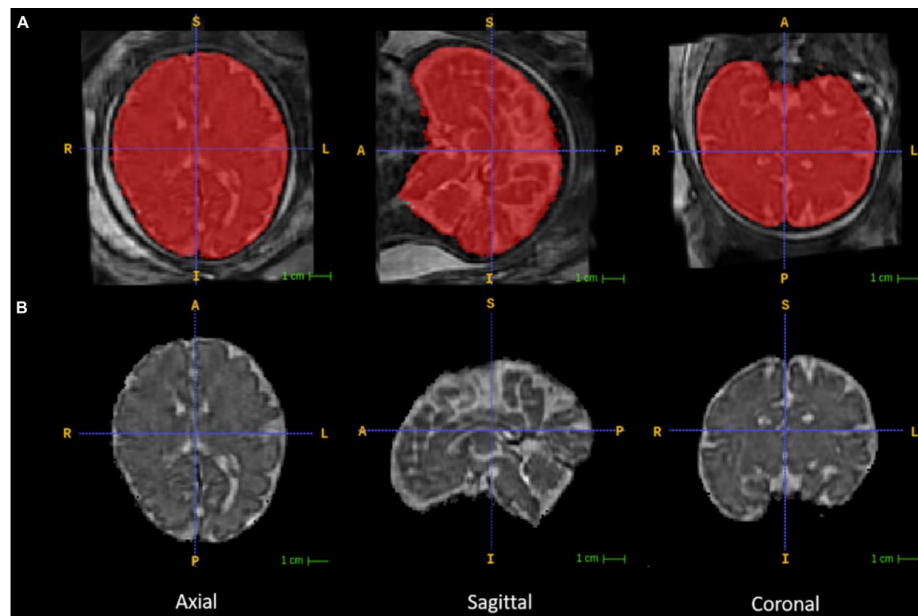
Registration-based subcortical segmentation

The average 36-week GA fetal brain atlas, including cerebellum and thalamus labels. The axial, sagittal, coronal, and 3D rendered views of the age-appropriate fetal brain atlas whereby deep brain tissues are color-coded.



Registration-based subcortical segmentation

Fetal brain segmentation. Row (A) includes the volumetrically reconstructed fetal brains in three planes. The red areas are the manually segmented fetal brain binary masks. Row (B) includes the orientation-corrected and skull-stripped (using the binary masks in red) fetal brain volumes in three planes





Registration-based subcortical segmentation

The applied FLIRT registration tool implemented the correlation ratio similarity metric for linear (affine) registration as the default parameters.

The ANTs registration tool used a mutual information (MI) similarity metric for linear (rigid and affine) registration and nonlinear (SyN) registration. The cross-correlation (CC) and MI similarity metrics, provided in the ANTs toolbox, which are both sufficient for intra-modality registration were used for rigid, affine, and SyN registration algorithms.

The whole fetal brain, cerebellum, and thalamus volumes were computed from the skull-stripped fetal brain masks and subcortical masks.



Registration-based subcortical segmentation

Cross-correlation (CC):

- Definition: Cross-correlation measures the similarity between two images by computing the degree of overlap between their intensity values at corresponding pixel locations.
- Calculation: Mathematically, it involves multiplying the intensities of corresponding pixels in the two images and summing up these products over all pixels. A higher cross-correlation value indicates a stronger similarity between the two images.

Mutual Information (MI):

- Definition: Mutual Information measures the amount of information shared between two images or image regions by quantifying their statistical dependency.
- Calculation: It computes the entropy of the joint intensity distribution of the two images and the entropies of their individual intensity distributions. MI is calculated as the difference between the sum of individual entropies and the joint entropy. A higher MI value indicates a greater similarity between the images.



Manual subcortical segmentation protocol

The left and right thalamus and cerebellum were delineated in all reconstructed T2- weighted images. The 3D reconstructed T2-weighted images were visualized and segmented using ITK SNAP. The displays provided simultaneous coronal, sagittal and axial views of the brain and created a 3D image of the thalamus and cerebellum.

Bilateral thalamus and cerebellum masks were created through the visual identification and tracing of these brain regions in each slice. Steps of the segmentation protocol:(first thalamus then cerebellum. Segmentations were based on the intensity differences between white and gray matter and the thalamus and cerebellum were presented in approximately 40 and 50 slices respectively)

1-Segmentation of the cerebellum and thalamus. Dependent on the participants and the resolution of the images, the rater segmenting the images manually composed segmentations through all three viewpoints (sagittal, coronal, and axial) to ensure that the masks were accurate in all viewpoints. The initially completed segmentations were verified in the other views, and any incorrectly identified areas were omitted and revised.

2-Inspection of the 3D surface. The segmented cerebellum and thalamus masks were represented in a 3D display through ITK-SNAP. The surface of the cerebellum and thalamus is expected to be smooth throughout, so any areas on the masks that protruded excessively were trimmed through a smoothing feature on ITK-SNAP. (Trimming the areas on the masks that protruded excessively with smoothing feature of ITK-SNAP)

3- Segmentation of left and right hemispheres. Once complete, cerebellum and thalamus masks were segmented into left and right hemispheres. Each mask was segmented and split into the left and right hemispheres by identifying the brain's midline. These segmentations were verified across all three viewpoints to ensure accuracy and to revise the original segmentations.(Each mask was segmented and split into the left and right hemispheres by identifying the brain's midline)



Protocol reliability testing

Three fetal MR images were randomly selected and re-segmented by the same rater to assess the reliability of the three-step manual segmentation protocol. The resegmentations of the left and right thalamus and cerebellum in the fetal MR images were performed 6 months after the original segmentations to ensure that the rater's memory would not unduly influence the results. This type of test-retest metric, intra-rater reliability, can be used as an upper bound metric to assess the accuracy of the segmentations of the thalamus and cerebellum. The protocol's reliability was measured using the Dice similarity metric, which evaluates the spatial and volumetric overlap of the original and resegmented label volumes.



Manually adjusting automatically generated masks from NiftyMIC

whole brain masks were manually segmented in all 25 fetal brain scans. The whole brain appeared in approximately 90 slices.

1-Automatic segmentation. Whole brain masks were generated automatically for each subject using NiftyMIC software.

2-Manual segmentation. Brain masks generated automatically through NiftyMIC were contrasted against the original brain scan for each subject on ITK-SNAP. Each mask was manually edited to ensure that the mask fit the image. Dependent on the subject and the clarity of the image, the individual segmenting the images manually worked through all three viewpoints (sagittal, coronal, and axial) to ensure that the masks were accurate in all viewpoints. The initially completed segmentations were verified in the other views, and any incorrectly identified areas were omitted and revised. Any area of the mask that protruded excessively outside the brain region was removed. Additionally, any areas of the brain that were not covered by the mask were filled in appropriately.

3-Segmentation of left and right hemispheres: Once the segmentations were complete, the whole brain masks were segmented into left and right hemispheres. Each mask was segmented and split into the corresponding hemisphere by identifying the midline of the brain.



Statistical analysis

The robustness of the entire automatic fetal deep brain structure segmentation workflow was tested by comparing the automatically segmented masks and manually segmented masks by calculating the DSCs of the common areas covered. $D = 2(A \cap B) / A + B$, where A and B represent the automatic and manual masks (0 no spatial overlap between automatic and manually segmentation results, 1 complete overlap). The masks for the left and right thalamus and cerebellum were combined.

The resulting DSCs were non-normally distributed, based on Shapiro–Wilk's tests (all, $p < 0.02$). \implies a nonparametric Friedman's test for repeated measures data was applied to the DSCs.

Results: Two-dimensional fetal brain segmentation and 3D volumetric reconstruction

The 2D fetal brain masks of the stacks of the original fetal brain MR images were automatically segmented using NiftyMIC in the axial, coronal, and sagittal image planes. For the NiftyMIC volumetric reconstruction algorithm to perform optimally, the 2D auto-masks were manually adjusted using ITK-SNAP for the over- and under-estimations of fetal brain tissue by the NiftyMIC segmentation algorithm. The volumetric reconstruction process was performed. Skull stripping and orientation tags correction were successfully applied to the reconstructed 3D volumes. Based on the skull-stripped automatically reconstructed 3D fetal brain MR images, manual segmentations of the thalamus and cerebellum on both left and right sides were successfully performed.

A segmented and volumetrically reconstructed fetal brain image using NiftyMIC. The original 2D slices of fetal MR images were automatically segmented and manually adjusted for fetal brain 2D masks. Then the 2D slices and 2D brain masks were reconstructed into 3D volumes and 3D masks with motion correction. This figure shows an example of the skull-stripped, orientation-adjusted 3D fetal brain volumes in axial, sagittal, coronal, and 3D-rendered views.

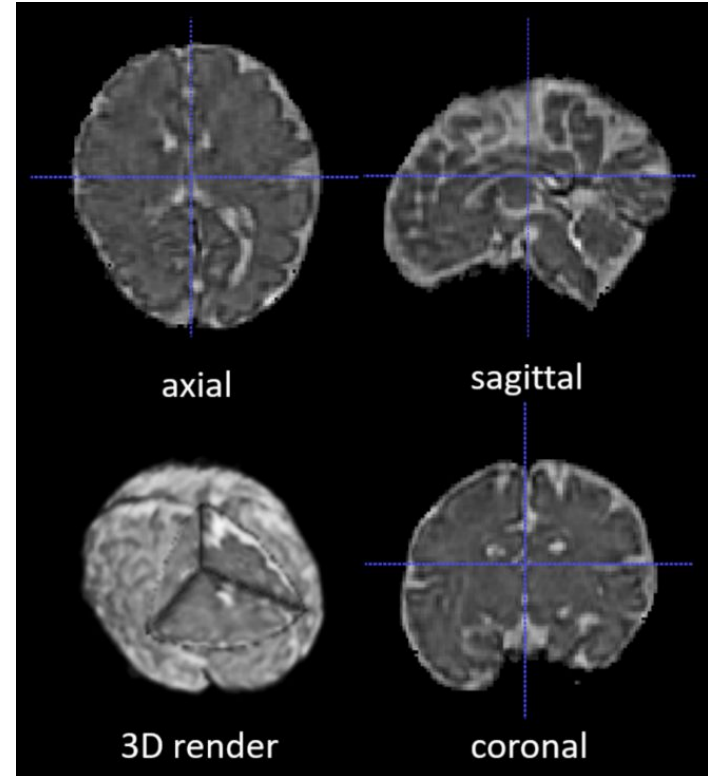


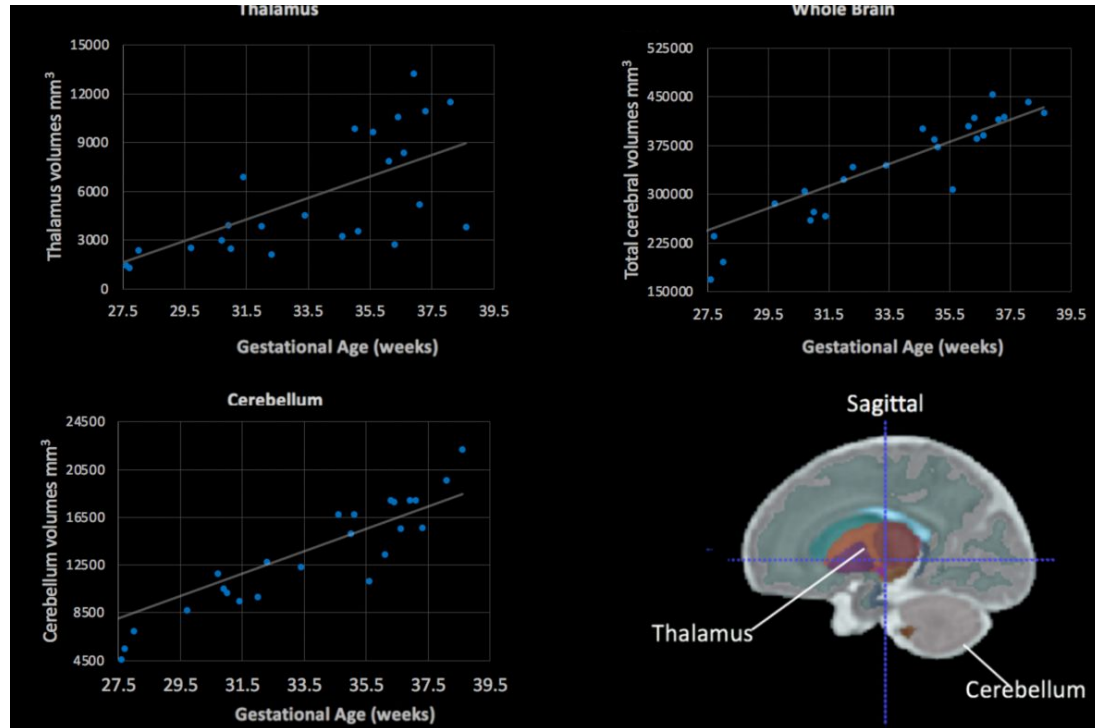


TABLE 2 Fetal brain volumes.

Characteristic	Total (<i>n</i> = 25)
Cerebellum, median volumes mm ³ (IQR)	13,365 (10,167–17,783)
Thalamus, median volumes mm ³ (IQR)	3,850 (2,714–8,381)
Whole brain, median volumes mm ³ (IQR)	373,186 (285,450–405,289)
GA, median weeks (IQR)	34.6 (30.9–36.4)

GA, Gestational age (weeks), IQR, interquartile range (25%ile–75%ile).

Of the 25 scans, the majority ($n = 21$) were completed on a 3T MRI and 4 were completed at 1.5T. None of the manually segmented volumes for the thalamus, cerebellum or total cerebral volumes differed based on the Tesla strength of the magnets when adjusting for gestational age (all, $p > 0.05$). The averaged left and right thalamus and cerebellum volumes were plotted against gestational age (Figure 5). All regions, the cerebellum ($r = 0.74$, $p < 0.001$), thalamus ($r = 0.7$, $p < 0.001$) and the total cerebral volumes ($r = 0.8$, $p < 0.001$) were positively associated with gestational age, indicative of larger volumes at older gestational ages.



Manual segmentation protocol validation: Intra-reliability test

The thalamus and cerebellum were re-segmented by a single rater (MM) to assess the consistency of the three-step manual segmentation protocol. The intra-reliability test results are listed in Table 3. The IQR of the median DSCs of cerebellar and thalamic segmentations were 0.78 and 0.6, respectively. The overall median DSC was 0.7

TABLE 3 Intra-reliability test – Dice similarity coefficients.

	Dice similarity coefficients
Cerebellum	0.78 (0.7–0.8)
Thalamus	0.6 (0.5–0.7)
Overall	0.7 (0.5–0.7)

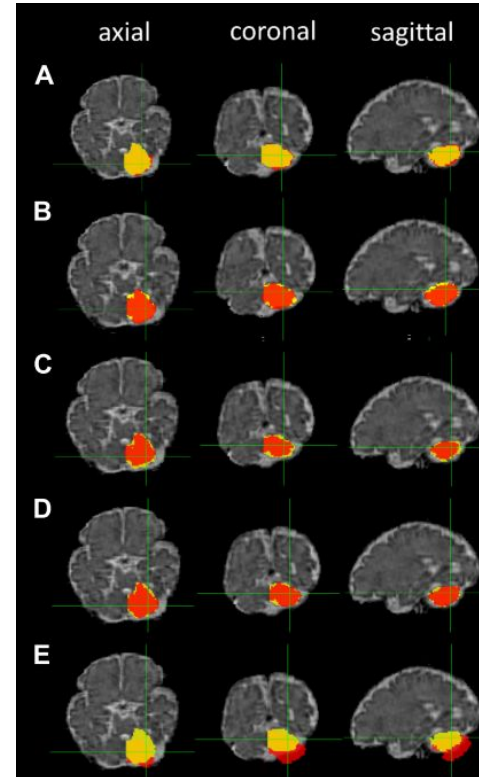
The median Dice similarity coefficients for cerebellar and thalamic segmentations, and both segmentations combined. IQR, interquartile range (25%ile–75%ile).

Registration-based segmentation reliability test: Comparisons of dice similarity coefficients

The ANTs- (5–10 h/dataset) and FLIRT-based (10 min/dataset) registrations of the 36-week GA fetal brain atlas into the native spaces of the individual fetal MR images were successfully processed in all participants. The median DSCs comparing the five image registration methods to the manual segmentation method were: (1) FLIRT linear registration (affine) using the correlation ratio similarity metric, (2) ANTs linear registration (rigid and affine) using the MI similarity metric (ANTs Lin MI), (3) ANTs linear registration using the CC similarity metric (ANTs Lin CC), (4) ANTs nonlinear registration (rigid, affine, and SyN) using the MI similarity metric (ANTs NL MI), and (5) ANTs nonlinear registration using the CC similarity metric (ANTs NL CC) for left and right cerebellum and thalamus segmentations. The cerebellar masks produced by the five registration methods using different similarity metrics are shown in Figure 6.

Registration-based segmentation reliability test: Comparisons of dice similarity coefficients

Cerebellar masks: registration-based segmentation (red) versus manual segmentation (yellow). The masks are shown in axial, coronal, and sagittal planes from left to right. Row (A) ANTs linear registration (MI); (B) ANTs linear registration (CC); (C) ANTs nonlinear registration (CC); (D) ANTs nonlinear registration (MI); and (E) FLIRT linear registration.



Registration-based segmentation reliability test: Comparisons of dice similarity coefficients

The median DSCs of the five registration methods for the cerebellum segmentations, thalamus segmentations, and both segmentations are listed in Table 4. Overall, the FLIRT linear registration resulted in non-optimal estimation with gross misalignment of the masks on the fetal MR image. The ANTs nonlinear registration (CC) had the highest median Dice similarity index. The ANTs non linear registration (MI) also demonstrated a very good performance.

The median DSCs of both subcortical segmentations revealed that ANTs NL CC and ANTS NL MI were high with the linear registrations being comparable, while those produced by FLIRT were the lowest

Registration method	Both	Cerebellum	Thalamus
FLIRT	0.54 (0.44–0.63)	0.62 (0.46–0.73)	0.52 (0.39–0.66)
ANTs Lin MI	0.70 (0.58–0.74)	0.80 (0.73–0.83)	0.59 (0.48–0.7)
ANTs Lin CC	0.72 (0.59–0.75)	0.80 (0.74–0.83)	0.61 (0.48–0.71)
ANTs NL MI	0.72 (0.63–0.76)	0.79 (0.75–0.83)	0.62 (0.49–0.68)
ANTs NL CC	0.74 (0.65–0.76)	0.79 (0.76–0.82)	0.65 (0.52–0.71)

The median Dice similarity coefficients and the interquartile ranges of the cerebellum, thalamus, and both subcortical segmentations using five registration methods compared to manual segmentations.

Registration-based segmentation reliability test: Comparisons of dice similarity coefficients

Subsequently, the DSCs produced by the linear and nonlinear registrations algorithms compared to the manual segmentations were examined for the left and right thalamic segmentations and were also significantly different ($n = 50$, $df = 4$, test statistic = 47.36, $p < 0.001$). Pairwise comparisons indicated slightly different results than seen for the cerebellar segmentations, whereby FLIRTbased registrations were associated with significantly different mean ranks compared to the ANTs-based nonlinear registration methods, including ANTs NL MI and NL CC, but also the ANTs Lin CC method (all $p < 0.002$; Table 6)

TABLE 6 *Post hoc* comparisons for mean ranks: fetal thalamic segmentations.

	Standard test statistic	<i>P</i> -value*
FLIRT – ANTs Lin MI	2.15	0.32
FLIRT – ANTs Lin CC	−3.67	0.002
FLIRT – ANTs NL MI	4.49	<0.001
FLIRT – ANTs NL CC	6.45	<0.001
ANTs Lin MI – ANTs Lin CC	−1.52	0.9
ANTs Lin MI – ANTs NL MI	2.34	0.19
ANTs Lin MI – ANTs NL CC	−4.30	<0.001
ANTs Lin CC – ANTs NL MI	0.82	0.1
ANTs Lin CC – ANTs NL CC	2.78	0.05
ANTs NL MI – ANTs NL CC	−1.96	0.5

Results of a Dunn's pairwise *post hoc* tests on the mean ranks. *Bonferroni corrected for multiple comparisons. Significant values are in bold.

Registration-based segmentation reliability test: Comparisons of dice similarity coefficients

The DSCs for the linear (i.e., ANTs rigid and affine, and FLIRT affine) and nonlinear (i.e., ANTs nonlinear with MI and CC similarity metrics) methods for the thalamus and cerebellum segmentations were compared using Friedman's Repeated Measure Analysis of Variance by Ranks. Upon comparison of the left and right cerebellar DSCs ($n = 50$), the calculated mean ranks were significantly different from one another ($df = 4$, test statistic = 100.84, $p < 0.001$). Post hoc pairwise comparisons revealed that the mean ranks were significantly different for the FLIRT-based registrations compared to the ANTs linear and nonlinear methods (all, $p < 0.001$; Table 5). Additionally, none of the mean ranks differed for any of the ANTs based registration methods (all $p > 0.88$).

TABLE 5 *Post hoc* comparisons of mean ranks: fetal cerebellar segmentations.

Sample 1–sample 2	Standard test statistic	P-value*
FLIRT – ANTs NL CC	6.77	<0.001
FLIRT – ANTs Lin MI	7.65	<0.001
FLIRT – ANTs NL MI	8.41	<0.001
FLIRT – ANTs Lin	–8.48	<0.001
ANTs NL CC – ANTs Lin MI	0.89	0.9
ANTs NL CC – ANTs NL MI	1.64	0.9
ANTs NL CC – ANTs Lin	–1.71	0.88
ANTs Lin MI – ANTs NL MI	0.76	0.9
ANTs Lin MI – ANTs Lin CC	–0.82	0.9
ANTs NL MI – ANTs Lin CC	–0.06	0.9

Results of a Dunn's pairwise *post hoc* tests on the mean ranks. *Bonferroni corrected for multiple comparisons. Significant values are in bold.

Registration-based segmentation reliability test: Comparisons of dice similarity coefficients

Comparison of the mean ranks indicated that ANTs NL CC performed significantly better than ANTs Lin MI ($p < 0.001$). We further compared the volumes extracted by the 5 registration methods relative to the manually segmented volumes. The extracted volumes for the cerebellum and thalamus based on the FLIRT and ANTs-based methods were subtracted from the manually segmented volumes. The differences in the volumes were then divided by the manually segmented volumes and the resulting values were converted to percentages (Figure 7). Overall, the cerebellar segmentations were more likely to be underestimated by ANTs-based methods. FLIRT-based registration of the thalamus and the cerebellum resulted in overestimation of the volumes.

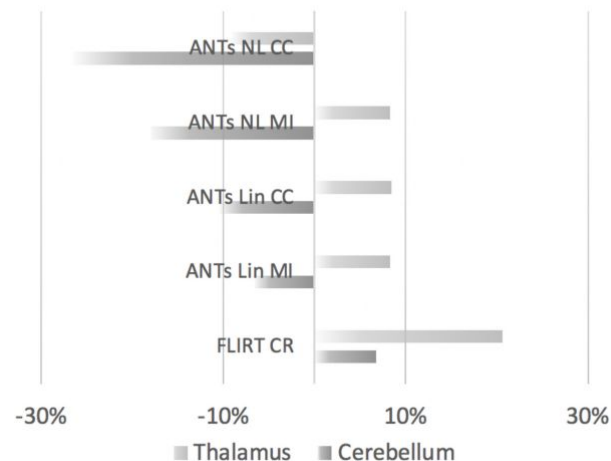



FIGURE 7

Comparisons of thalamic and cerebellar volumes produced by FLIRT- or ANTs-based methods relative to manually segmented volumes. The overlap (positive values indicate overestimation, negative values indicate underestimation) is displayed according to the registration methods from top to bottom. Top to bottom: ANTs nonlinear registration (CC, cross-correlation); ANTs nonlinear registration (MI, mutual information); ANTs linear registration (CC); ANTs linear registration (MI); and FLIRT linear registration.

Conclusion



Deformable registration methods(ANTs using the well-known SyN (symmetric image normalization) method) provided the most accurate results relative to manual segmentation. Overall, this semi-automatic subcortical segmentation method provides reliable performance to segment subcortical volumes in fetal MR images. This method reduces the costs of manual segmentation, facilitating the measurement of typical and atypical fetal brain development.



Thanks!