Observing the origins of human brain development: Automated processing of fetal fMRI

Saige Rutherford [a*], Pascal Sturmfels [b], Mike Angstadt [a], Jasmine Hect [c], Jenna Wiens [b], Marion I. van den Heuvel [d], Dustin Scheinost [e,f,g], Moriah Thomason [h,i**], Chandra Sripada [a**]

[a] Department of Psychiatry, University of Michigan, Ann Arbor, MI
[b] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI
[c] Department of Psychology, Wayne State University, Detroit, MI
[d] Department of Cognitive Neuropsychology, University of Tilburg, Tilburg, The Netherlands
[e] Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT
[f] Department of Statistics and Data Science, Yale University, New Haven, CT
[g] Child Study Center, Yale School of Medicine, New Haven, CT
[h] Department of Child and Adolescent Psychiatry, New York University School of Medicine, New York, NY
[i] Department of Population Health, New York University School of Medicine, New York, NY

*Corresponding author
Email address: saruther@umich.edu
**Contributed equally to this work

Abstract

Fetal resting-state functional magnetic resonance imaging (rs-fMRI) has emerged as a critical new approach for characterizing brain network development before birth. Despite rapid and widespread growth of this approach, at present we lack neuroimaging analysis pipelines suited to address the unique challenges inherent in this data type. Here, we present a comprehensive fetal data processing pipeline with provision of full code, and solve the most challenging processing step, rapid and accurate isolation of the fetal brain from surrounding tissue across thousands of non-stationary 3D brain images. Leveraging our library of more than 1,000 manually traced fetal fMRI images, we trained a Convolutional Neural Network (CNN) that achieved excellent accuracy (>90%) across two held-out test sets from separate populations. Furthermore, we unite the auto-masking model with a full processing pipeline and report methodology, performance, and comparison to available alternatives. This work represents the first open source, automated, start-to-finish processing pipeline for fetal fMRI data.

Keywords: fetal, fMRI, functional imaging, brain segmentation, convolutional neural network, automated preprocessing, open-source software

1. Main

Resting-state functional magnetic resonance imaging (rs-fMRI) has emerged as a powerful tool for studying development of the brain's network architecture. In recent years, this methodology has been applied to the human brain *in utero*, opening a window into a period of functional development that was otherwise inaccessible. Studying fetal fMRI has the potential to illuminate the nature and manner in which the brain's network architecture is initially assembled, affording powerful new insights into the neurodevelopmental origins of major mental illness (Jakab et al. 2014, 2015, Schöpf et al., 2012, Thomason et al., 2017, van den Heuval et al., 2018). Despite this potential, progress has been slow due, in part, to the lack of image analysis tools tailored for fetal imaging data. Though many tools and software packages exist for fMRI analysis, these tools were designed with adult and child data in mind and encounter specific problems when applied to fetal functional data. In particular, these tools are not set up to handle encasement of the head inside a heterogeneous tissue compartment, high degree of variation in image properties across subjects, and the extensive position adjustments made by the fetus during a typical scanning session.

The majority of progress in fetal MRI methodology to date has occurred with respect to fetal *structural* MRI, particularly in anatomical (primarily T2 HASTE; Half-Fourier Acquisition Single-shot Turbo spin Echo imaging) and diffusion tensor imaging (DTI) (Benkarim et al., Biegon et al., 2014). Advances have been made in inter-slice motion correction and volume reconstruction (Gholipour et al. 2017, Fogtmann et al., 2014, Studholme et al., 2011), mapping structural connectivity (Ouyang, M. et al., 2018, Qiu et al., 2015, Ouyang, A et al., 2015, Takahashi et al., 2012, Song et al., 2017, Huang et al., 2017), comparing different MRI signals (Seshamani et al., 2015) and segmentation of the brain from T2-weighted anatomical volumes (Klinder et al., 2015, Ison et al., 2012, Makropoulos et al., 2017, Link et al., 2017, Serag et al., 2017, Rajchl et al., 2016, Tourbier et al., 2017, Salehi et al., 2017). These strategies have also made it possible to use sparse acquisition sequences, which alleviate movement concerns (Seshamani et al., 2014, Serag et al. 2017), and enable more sophisticated analytic approaches, such as morphometric (Gholipour et al. 2017, Studholme et al., 2011, Kuklisova-Murgasova et al. 2011, Serag et al., 2012, Shi et al., 2011), cortical folding

(Wright et al., 2014), and cytoarchitectural examinations (Miller et al., 2014). In contrast, papers providing possible solutions for the analysis of fetal *functional* MRI data are markedly few (Seshamani et al., 2015, Scheinost et al., 2018).

Challenges associated with analysis of fetal fMRI have been discussed in a growing number of studies (Thomason et al., 2014, 2015, 2017, 2018, van den Heuval et al., 2018, Schuler et al., 2018) and reviews (Anderson et al., 2013, van den Heuval et al., 2016, Robinson et al., 2018, Vasung et al., 2018). These works have focused on image characteristics: motion, size of the fetal brain, susceptibility artifacts introduced by surrounding maternal tissues, and physiological noise of both mother and fetus. Previous work has highlighted important areas for development, but to our knowledge, no one has performed a careful examination of strategies for processing fetal fMRI. Moreover, no fetal fMRI pipelines have been developed and released in the open science framework.

The most time-consuming step in preprocessing fetal fMRI is differentiation of the fetal brain from the surrounding maternal compartment at each acquisition time point. This is achieved by generation of an exemplar mask that marks all in-brain voxels. This mask is critical for the entire preprocessing pipeline and for subsequent activation and/or connectivity analyses. Tools developed to segment the adult brain, such as the Brain Extraction Tool (BET) from FSL (Jenkinson et al., 2002) and 3dSkullstrip from AFNI (Cox et al., 1996) are not effective in generating exemplar masks in fetal imaging because the surrounding tissue is more complex and the fetal brain is not in a standard orientation, making it more challenging to identify. As a result, previous studies have relied on manual generation of brain masks (Thomason et al., 2013, 2014, 2015, 2017, van den Heuval et al., 2018). While manual methods are tedious and time consuming, to date, they have been necessary to achieve acceptable standards.

Here, we present a novel automated approach to the problem of fetal brain segmentation from surrounding tissue. Leveraging a large corpus of manually traced human fetal fMRI masks, we trained a convolutional neural network (CNN) to replace this labor-intensive preprocessing step. CNN's are a powerful tool for effectively identifying complex, non-linear patterns in spatially structured high-dimensional datasets (LeCun et al., 1989). They are increasingly utilized in image processing applications in both medical and non-medical settings (Egmont-Petersen et al., 2002,

Zeiler et al., 2014, Ronneberger et al., 2016). In the context of fetal brain segmentation, prior work has investigated the application of CNN's to segment fetal structural T2-weighted volumes (Rajchl et al., 2016, Salehi et al., 2017). These models, however, were developed to segment the fetal brain from anatomical images, and do not translate to functional time series data. Compared to structural data, functional data is significantly lower resolution and, due to movement, requires a larger quantity of individual segmentations. Here, we extend prior work by developing and validating new tools for automatically segmenting the fetal brain from functional data in the largest sample of fetal fMRI data published to date.

In addition to segmentation of the fetal brain from surrounding maternal tissue, fetal imaging preprocessing requires a number of additional steps: motion denoising, realignment of volumes within a time series, and group-level normalization to a fetal template. All of these steps are challenging because frame-to-frame displacement is elevated in fetal studies, and the fetal brain is typically not in a single standard orientation. To address these additional issues, we integrate our novel segmentation tool into a new open source, automated fetal fMRI preprocessing pipeline. This pipeline inputs raw fetal fMRI data and outputs fully preprocessed, ready to be analyzed data. All code discussed in this paper, along with a protocol, is available on GitHub (https://github.com/saigerutherford/fetal-code). This pipeline embodies an initial set of best practices for fetal fMRI processing, and associated protocols are expected to evolve over time through improvements by the user community in response to new knowledge and innovations in the field.

2. Results

Our CNN auto-mask model learned to accurately discriminate fetal brain from surrounding structures in fetal brain fMRI images. We evaluated the model on two held-out test sets. Applied to the Wayne State University (WSU – 48 subjects) and Yale (10 subjects) test cohorts, the model achieved a per-volume average dice coefficient of 0.92 and 0.91, respectively. The CNN's performance in terms of dice coefficient, sensitivity, specificity, and per-pixel accuracy across both test sets is summarized in Table 1. Figures

1 and 2 provide examples of agreement between manual and auto-masks in both test sets.

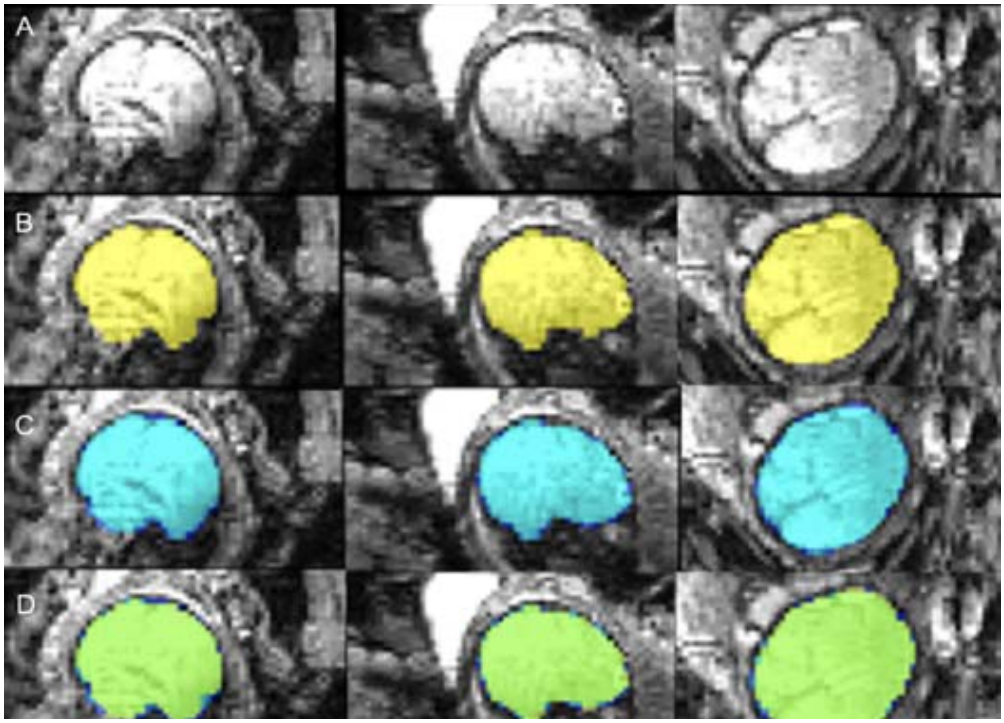

Figure 1. Comparison of manual and automated masks in WSU cohort. A) Raw volume; B) Hand drawn mask; C) Auto mask; D) Conjunction of hand drawn (yellow) and auto (blue) masks, overlap between hand and auto masks shown in green. *Data collected in Detroit, MI at Wayne State University*

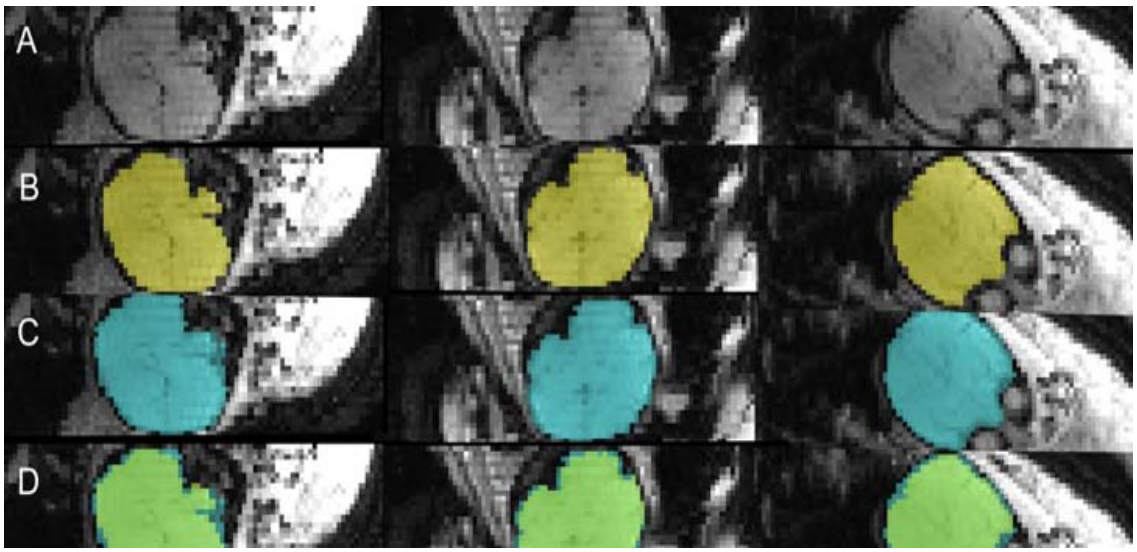

Figure 2. Comparison of manual and automated masks in Yale cohort. These results demonstrate the auto-mask model's ability to transfer to data collected in different populations and/or scanners than the original data used to train the model. A) Raw volume; B) Hand drawn mask; C) Auto mask; D) Conjunction of hand drawn (yellow) and auto (blue) masks, overlap between hand and auto masks shown in green. *Data collected in New Haven, CT at Yale University.*

Examination of the effect of fetal age on performance of the algorithm revealed a significant positive association between dice coefficient and gestational age, r=0.54, p=4.84x10^{-5}. This effect is demonstrated in Figure 3. This relationship may result from older fetuses having larger brain volumes that intrinsically have higher effective image resolution. We also found that significant aliasing, particularly phase wrap-around negatively impacted auto-mask performance, and that the algorithm also performed more poorly for images in which the brain was far off center of the imaging space. Performance improved with an increasing number of augmentations in the training dataset.

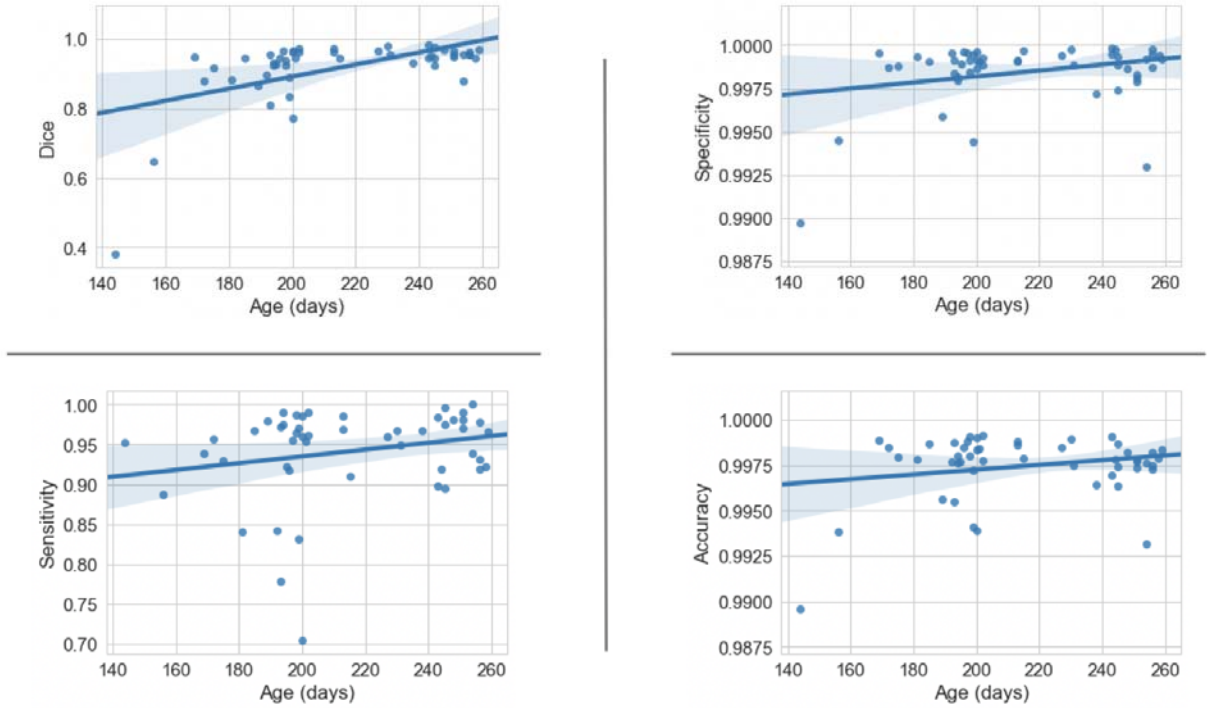

Figure 3. Evaluation of auto-masking model. The relationships between fetal gestational age at scan and performance of auto-masking in the WSU test set (48 subjects, 211 volumes). We calculated the Dice coefficient, specificity, sensitivity, and per-pixel accuracy on a per-volume basis. However, the values shown here are on a per-subject basis in order to examine the relationships with age. There were significant correlations between age and performance evaluated by the Dice coefficient (r=0.54, p=4.8x10^{-5}) and specificity (r=0.28, p=0.05). The relationships between fetal gestational age and sensitivity (r=0.22, p=0.12) and per-pixel accuracy (r=0.23, p=0.11) were not statistically significant.

Table 1.
Performance of the CNN on the held-out test sets, with 95% bootstrapped confidence intervals.

	WSU	Yale
Dice	0.92 (0.90, 0.93)	0.91 (0.89, 0.93)
Sensitivity	0.94 (0.93, 0.95)	0.94 (0.90, 0.96)
Specificity	0.99 (0.99, 0.99)	0.99 (0.99, 0.99)
Per-pixel Accuracy	0.99 (0.99, 0.99)	0.99 (0.99, 0.99)

As a point of reference for the dice coefficient achieved by our method, we performed an additional analysis using BET, the most widely used alternative for brain auto-extraction. Applied to the same test set data, BET performed significantly worse, achieving an average dice coefficient of 0.28. An example BET mask is shown in figure 4. This image highlights that areas of the maternal compartment have high contrast boundaries and varied image intensity creating serious challenges for standard masking routines and severely compromise performance.

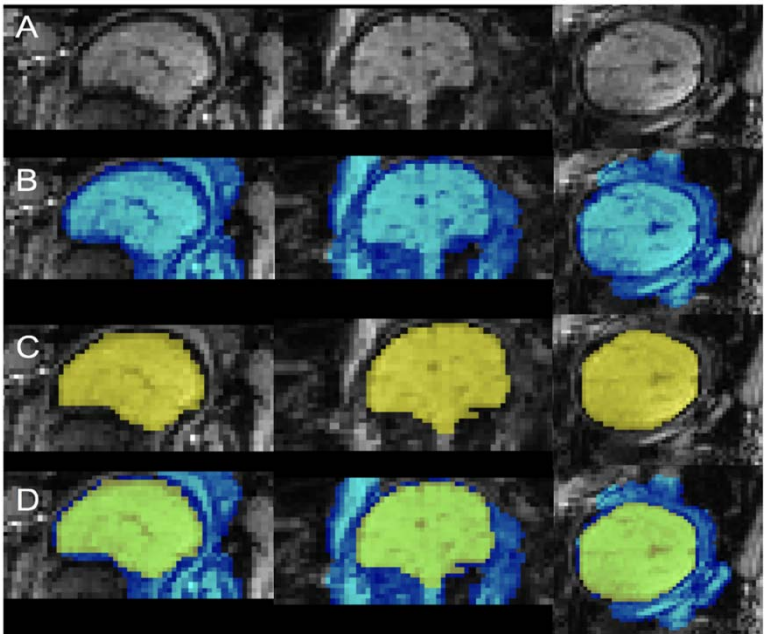

Figure 4. Limitations of the existing toolkit for automated brain masking. A) Raw volume; B) BET mask; C) Hand drawn mask; D) Conjunction of hand drawn (yellow) and BET (blue) masks. The BET masks do not properly capture the fetal brain's boundary.

An often-noted property of deep learning models is their ability to substantially surpass human speed in completing complex tasks. Our auto-masking model illustrates this acceleration. The training time refers to the wall-clock time it took our CNN model to converge to a set of weights that minimize the dice coefficient on the validation set. Training was stopped after signs of overfitting were observed, that is, performance on the validation set was no longer increasing. The total training time of the model was 3 hours and 46 minutes on a GeForce GTX 1080 Ti GPU. Testing time refers to the time it takes to run the CreateMask.py script in order to load the input raw volume (nifti file type to numpy array conversion) and output a predicted auto-mask. Additionally, we report the time to create auto-masks for all volumes within a subject's time series. As some potential users of these preprocessing methods may not have access to GPU computing resources, we report testing time in GPU as well as CPU environments. These testing times are summarized in Table 2.

Table 2.
Average computing times of all preprocessing steps. The time to run the full pipeline is for a single subject with 360 volumes. Testing GPU was a NVIDIA Quadro P6000.

	GPU	CPU
Auto-mask a single volume	0.2s	2.5s
Auto-mask an entire time series	1.2m	15m
Realignment	-	5m
Normalization	-	7m
Time to run full pipeline	13.2m	27m

An additional benefit of our approach for auto-masking all individual volumes is that multiple realignment strategies are now possible. After the fetal brain has been extracted, the time series data can enter more typical preprocessing steps for which child/adult tools have been developed. The main difference when applying these tools is that the fetal brain is commonly in a non-standard orientation, and fetal data exhibits substantially increased head motion.

To address the first challenge, we incorporated realignment using MCFLIRT (Jenkinson et al., 2002) into the pipeline. This method uses the preceding volume to provide an initial realignment estimate for the current volume. The idea behind this choice is that there is likely to be high spatial displacement between volumes 10 and 100 because they were collected 180 seconds apart, but less displacement between volumes 10 and 11 collected just 2 seconds apart.

With regard to the problem of elevated head motion, errors introduced by movement cannot, at present, be fully corrected in fMRI time series. This fact necessitates application of stringent criteria for retaining only low-motion volumes to ensure data integrity, which is the approach taken by most studies to date (Jakab et al. 2014, 2015, Schöpf et al., 2012, Thomason et al., 2017, van den Heuvel et al., 2018). We recommend two criteria for retaining low motion volumes: 1) framewise displacement less than 0.5mm, and 2) number of consecutive volumes reaching those criteria must be 10 or more. The latter rule reduces the number of breaks in the time series introduced by the data elimination scheme. These criteria await systematic testing against alternative schemes. The pipeline introduced here is built flexibly so that these thresholds can be modified, allowing users to apply alternative motion control schemes. An overview of the entire preprocessing stream is provided in Figure 5.

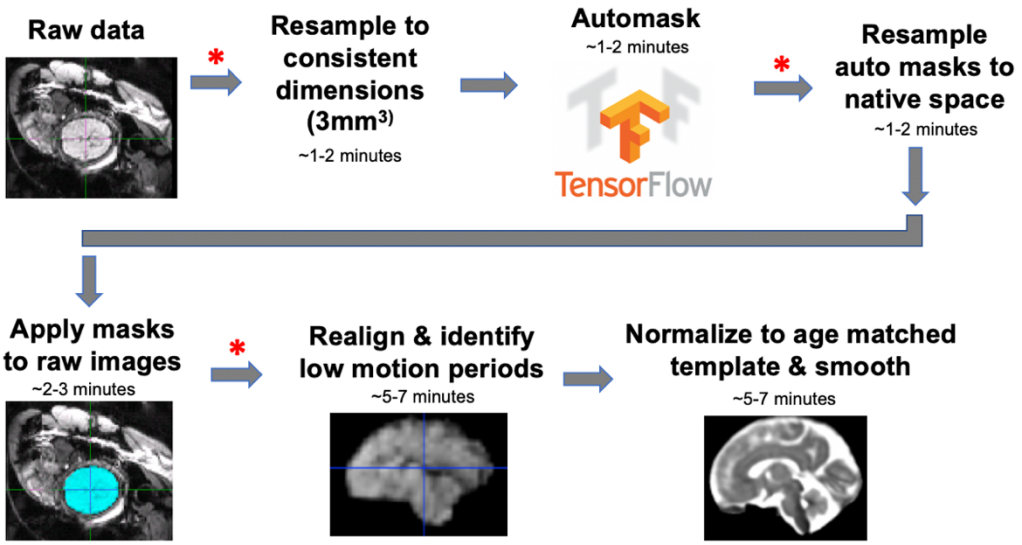

Figure 5. Overview of proposed processing pipeline. All steps in the proposed preprocessing stream are shown, with red asterisk representing where visually quality checking data is recommended. This workflow has been implemented in Python and can also be run as shell scripts from the command line.

3. Discussion

Fetal functional MRI is an emerging field with great potential improve understanding of human brain development and origins of mental illnesses. The number of papers published in this area has seen a 5-fold increase since 2010. Methodologies for processing these complex data sets, however, have not kept pace, and the absence of a standard publicly available processing pipeline for these data has been especially notable. Here we address this gap and present a comprehensive fetal fMRI data preprocessing pipeline that was built by pairing a set of more than 1,000 hand-drawn fetal 3D brain masks with a powerful deep learning CNN algorithm. Pipeline code, test data sets, and documentation are made available through GitHub (https://github.com/saigerutherford/fetal-code). It is hoped that release of an easy-to-use, efficient, validated preprocessing pipeline will reduce barriers for new labs to enter this area, while also providing experienced labs opportunities for further optimization of their independently developed approaches.

We have built a suite of tools that automate analysis from raw fetal fMRI time series data to fully preprocessed individual subject data. Steps covered between raw and fully preprocessed data include brain segmentation, reorientation, realignment, normalization, and smoothing, all of which are tailored to the unique properties of fetal fMRI brain data. At the core of this approach is an adaptation of a CNN that was used to train a solution to the most cumbersome processing step, localization and extraction of the fetal brain from surrounding tissue for each volume of a functional time series. This is a necessary step in processing human fetal fMRI data and until now, it has been a rate-limiting factor in accurate, automated processing of fetal fMRI BOLD timeseries data.

Recent pioneering work by Salehi and colleagues (2017) established that deep learning approaches can be effective in fetal *structural* brain segmentation. However, fetal *functional* imaging presents a distinct set of constraints, and therefore requires a different solution. In particular, the inherently lower resolution and contrast of functional timeseries data, and the 4-dimensional nature of the data (~360 3D volumes per subject), make this a more challenging problem. Here we applied CNN methods to the largest fetal fMRI data set reported to date, 207 fetuses, and derived a novel method

for accurate and reliable segmentation of the fetal brain from surrounding maternal tissues within a fraction of a second, 92% accuracy in 0.2 seconds. These encouraging results are partially attributable to the large set of manually traced human fetal fMRI masks used to train the CNN, consistent with studies showing the correlation between CNN performance and the size of the training data (Cho et al., 2015, Verghese et al., 2018).

The brain segmentation CNN exhibited signs of strong generalizability, which is particularly important given known biases in CNN models trained on imaging datasets from a single site/population (Tommasi et al., 2015, Zech et al., 2018). First, the trained CNN correctly classified data at two held out sets with performance, 91% and 92%, very similar to the training data. This finding suggests that the CNN is robust to variations in experimental procedures, scanner settings, and populations studied. In addition, the training images were drawn from a wide fetal age-range, which should also enhance generalizability across fetal samples encompassing different ages.

Our aim in this project goes beyond brain segmentation; we sought to construct a full processing pipeline that is standardized but flexible, and readily deployable across multiple data sources. Thus, our comprehensive pipeline begins with a core segmentation step followed by masking, reorientation, frame to frame alignment, normalization to a user-defined template, and smoothing. The user is referred to publicly available multi-age fetal brain templates (Serag et al., 2012), and is able to easily configure the tool to modify or eliminate steps. The open construction of the tool will allow incorporation of future processing advances, such as such as surface-based registration and additive motion correction strategies. Finally, the pipeline uses Nipype and BIDS formats as these are accepted best standards for facilitating transparency and harmonization in neuroimaging data processing (Gorgolewski et al., 2011 and 2016). Of note, large scale, often multi-center, projects are becoming the new norm, and these require validated, standardized processing pipelines of the kind that we have developed. The Developing Human Connectome Project provides just one example of a large-scale study that includes a fetal functional MRI component (Garcia et al., 2018, Makropoulos et al., 2018), and many more large-scale fetal fMRI initiatives will likely emerge in the coming years.

Our work has several limitations. First, deep learning methods perform classification in high dimensional space, and consequently, results can be a “black box” with little opportunity for interpretation of axes (Cabitza et al., 2017). However, this limitation should be viewed in the context of our goal: to automatically perform a task that can take trained individuals many hours to perform manually. As such, we are less interested in understanding the mechanisms of computer-based brain masking, and instead focus on algorithm performance.

Another limitation is that we use direct warping, or normalization, of functional data to a group-averaged anatomical template. It is not clear that alternatives would improve registration significantly, but one might expect registration to subject-specific anatomy, then to template space, to be a preferred approach. The challenge of this approach is that obtaining high-quality subject-specific high-resolution anatomical images presents a different set of challenges that have been addressed elsewhere (Studholme et al., 2015). Ground-breaking studies of fetal anatomical development (Nunes et al., 2018) demonstrate that even when trained experts apply the most advanced techniques to these data, there is still significant data loss and image blurring where motion effects, image artifacts, or lack of tissue contrast compromise data quality. This example extends to other parts of the pipeline, where alternative optimized preprocessing strategies could be used. However, the objective for this work is not to serve as a final fetal fMRI data processing endpoint, but as a backbone upon which further development can follow.

A final limitation is that ours is not a fully automated pipeline as it requires human supervision and quality checking at several stages, which in turn requires a certain quantity of time and level of expertise from the human supervisor. Fortunately, however, the level of involvement, and associated expertise required, required is limited, and includes looking for overt errors when viewing processed images as a continuous movie, which takes approximately 1-5 minutes per functional run. Assuming the full time for running this pipeline is 15-30 minutes, including human effort, this is a 60-fold time reduction over prior methods, with manual tracing in particular requiring extensive time and substantial expertise (van den Heuvel et al., 2018, Thomason et al., 2017).

In sum, in this work, we leverage CNN methods to address the challenging brain segmentation problem in fetal fMRI, and we present the first complete open-source solution to processing raw fetal fMRI timeseries data.

4. Methods

4.1 Project Background

Primary data used for pipeline development were acquired at Wayne State University School of Medicine during the course of projects supported by National Institutes of Health (NIH) awards MH110793 and ES026022. These projects aimed to characterize the development of functional neural systems beginning *in utero* in relation to prenatal exposures and future neurobehavioral development. Participants were recruited from obstetric clinics located within the Detroit Medical Center. A physician with a clinical relationship to eligible patients initiated contact. Eligible participants were at least 18 years of age, assessed as having uncomplicated, singleton pregnancies, and had no contraindications for MRI. Those expressing interest in participating were then introduced to a member of the research team, who explained study procedures and answered patient questions. All participants provided written informed consent before undergoing MRI examination. The study protocol was approved by the Human Investigation Committee of Wayne State University.

4.2 Participants and Data

Resting-state functional MRI was obtained from two cohorts, Wayne State University (WSU) and Yale University. WSU cohort consists of 197 fetuses (gestational age 24-39 weeks, M=30.9, SD=4.2). Twenty-one of these fetuses were scanned at two time points *in utero*. Both time points are included in this study, however, they are counted as a single subject. WSU fetal MR examinations were performed on a Siemens Verio 3T scanner using an abdominal 4-Channel Flex Coil. Scanning protocols have evolved since the inception of the project in 2012. The majority of data were acquired using echo-planar sequence (TR/TE: 2000/30; 4mm slice thickness, axial, interleaved ascending slice order, 360 volumes) (See Supplementary Materials for all scan parameters). Multi-echo resting-state sequences were also collected in a portion of these

subjects (TR/TEs: 2000/18,34,50). The Yale University cohort contains 10 fetuses scanned twice longitudinally (gestational ages 30-36 weeks, M=32.7, SD=1.9). The Yale scanner was a Siemens Skyra 3T using a 32-channel abdominal coil (TR/TE:2000/30; 3mm slices, 32 slices parallel to the bi-commissural plane, 150 volumes).

Due to lack of tools for automated segmentation of the fetal brain, research personnel were trained to manually draw fetal brain masks using BrainSuite software (Shattuck et al., 2002). In line with prior work, in the present analysis manually generated brain masks are used to judge the accuracy of automated segmentation methods.

4.3 Auto masking

4.3.1 Experimental Pipeline

WSU data were randomly separated at the subject level into training, validation, and test sets with 129, 20, and 48 subjects (855, 102, and 211 volumes) respectively. The training set was used to optimize the model. The validation set was used to gauge the generalization performance of the network during training and to determine when to stop training. The test set was held out and used only after training to evaluate the performance of the model.

For those interested in using this model on unlabeled data, we re-trained the CNN a second time on both the training and held-out test data, and used the same validation set for determining when to stop training. We re-trained on as much data as possible to maximize the performance of the model on unlabeled data.

4.3.2 Data Preprocessing

Minimal preprocessing for data used to train and validate the auto-mask model included removing the image orientation from the image header, resampling and zero padding the images to consistent voxel sizes (3.5 x 3.5 x 3.5 mm) and dimensions (96 x 96 x 37). All nifti volumes were converted to NumPy arrays prior to training.

4.3.3 Network Architecture

The U-Net style CNN network architecture (Ronneberger et al., 2016) implemented in this pipeline was adapted from prior work by Salehi *et al.* (2017). The architecture features repeated blocks of 3x3 convolutions followed by the ReLu activation function assembled into a contracting path, followed by an expanding path. In the contracting path every second convolution is followed by a 2x2 max pooling operation. In the expanding path, every second convolution is followed by a 2x2 upsampling operation using nearest-neighbor interpolation. Every other feature map in the contracting path is concatenated along the depth dimension to the corresponding map in the expanding path, which helps the network learn the appropriate location of the output mask. The final layer is convolved with two 1x1 filters to produce an output mask with channels equal to the number of output classes.

The network separates 3D image volumes into 2D axial slices and operates on each slice independently. We chose to implement a 2D rather than 3D network in order to reduce computational costs. This model includes steps for converting raw NIFTI images into a format readable by the network, and steps for converting the output of the network into a NIFTI-formatted, 3D brain mask.

The model was implemented using Tensorflow (version 1.4.1). Training and testing of the network were performed using a GPU, but CPU testing times are were also evaluated to provide an additional point of reference.

4.3.4 Training Procedures

During training, the weights in a CNN are minimized with respect to a loss function that determines how well the network is learning from the training data. We optimized our network with respect to per-pixel cross entropy, with weights determined using the Adam Optimizer (Kingma et al., 2014). Adam is a first-order gradient method that updates the weights adaptively based on previous and current gradients. Even using an adaptive optimizer, we found that using a learning rate decay improved performance. The initial learning rate was set to 0.0001 with exponential decay rate of 0.9, applied every 10,000 batches. The model was trained until performance no longer improved on the validation set. In addition, we augmented the 2D axial slices in the training data through 90-degree rotations and horizontal and vertical flips. These augmentations capture the non-standard orientation of the brain in fetal volumes.

4.3.5 Evaluation

The evaluation process was performed over multiple steps. First, we evaluated our network's ability to mask the fetal brain using the dice similarity coefficient, D, defined as follows:

$$D = \frac{2\,|A \cap G|}{|A|+|G|}$$

where **A** refers to the auto-mask and **G** to the ground truth hand drawn mask. Intuitively, the dice coefficient measures the percent overlap between two regions: the predicted brain region and the true brain region. It is defined between 0 and 1, where 0 means there is no overlap between the two regions, and 1 means the two regions are identical. We also report the sensitivity, specificity, and per-pixel accuracy of our network on the WSU and Yale held-out test sets. In addition, to aid in evaluation of obtained dice values, we performed a secondary analysis to demonstrate current methods of adult brain extraction perform poorly when applied to fetal data. We used an alternative brain auto-masking technique, BET implemented in FSL, enabling comparison of approach efficacy. Of note, dice coefficients can be improved by separating testing data into challenging versus non-challenging images (Salehi et al., 2017), but that approach is not favored and not used here as this diminishes the representativeness of estimates when applied across complex and varied data sets.

After training, we calculated dice coefficients for all 1,168 auto-masks, though we report values only for volumes in the test data, as performance within the train and validation datasets does not reflect model performance on new data. To demonstrate the transferability of our model, we tested the CNN on an additional, outside fetal functional dataset (referred to as the Yale test set) collected in a separate population and on a different scanner. The Yale test set was comprised of 73 volumes from 10 unique subjects.

Once fits were determined using dice coefficients, several parameters with the potential to influence goodness of fit were tested in turn. First, we examined the relationship between the dice coefficient and gestational age. Next, we examined the effect of varying the number of augmentations in the training dataset. Finally, we evaluated whether image artifacts and/or position of the brain in the center of imaging

space influencing fit by qualitatively examined images with dice coefficients falling below 0.8.

4.4 Application of auto-masks

After auto-masks were generated for all available data, summarized in table 3, visual inspection was used to confirm accuracy and quality, and a pass/fail scale was applied. Approved auto-masks were output with spatial probability estimates, wherein voxel values equal to one correspond to highest probability of being brain. Failed auto-masks were discarded. Probability map brain masks were then clustered, thresholded, and binarized. These steps are taken in order to discard small, non-brain clusters that may have been included in the probability map brain mask. Binarized masks were then resampled back into subject native space where they were applied to the native image using a multiplier, resulting in segmented brain volumes corresponding to each fetal fMRI data timepoint.

Table 3.

Summary of all functional data collected at Wayne State University and preprocessed using the proposed automated preprocessing pipeline. The amount of time spent creating brain masks for all available data by automated methods is compared to the amount of time it would take to manually create brain masks.

Total number of subjects	248
Number of subjects scanned longitudinally	163
Total number of visits	411
Total number of runs	777
Total number of masked volumes	190,056
Time to auto-mask all volumes	10.6 hours
Time to manually mask all volumes	22 years

4.5 Motion quantification and de-noising

In prior studies by our group, a reference frame from each quiescent period was chosen to be masked. The mask would then be applied to every volume within the low movement period, not only the volume it was drawn on. Due to the time-consuming nature of manual masking, it was not feasible to mask every volume. A central goal of auto-masking is to be able to mask a full time-series in a fraction of the time it takes to manually draw a mask for a single volume. Realigning all volumes within a time series aids in automatically identifying low movement periods that are usable for further activation or connectivity analyses. Volumes in low movement periods are identified by the realignment parameters that are output as a text file and can also be visualized from the saved realignment plots. Typical realignment of fMRI data is done on full time series using the middle volume (in time) as a reference volume. Due to notably high movement across a fetal time series, using a single reference volume for realignment may not be an optimal approach. We selected the MCFLIRT FSL realignment tool (Jenkinson et al., 2002). While this tool still uses a single reference volume, it estimates a linear transformation between volume n and the reference volume and then uses this transform as the starting point to estimate the n+1 to reference transform. These transformation matrices are applied to each volume of the full time series to produce a new data set comprised of realigned volumes. This step also produces a text file and plot that summarize the six rigid-body realignment parameters across time, which can be subsequently used in identification of motion outliers and/or motion censoring in later processing stages. Here, we applied the fsl_motion_outliers routine as a data-driven means of defining periods of high and low fetal movement.

4.6 Preprocessing Protocol

After masking and realignment, time-series data are converted into a 4-dimensional file, moved into group template space (for multi-subject averaging) using linear warping, and spatially smoothed. Flexibility is built into the pipeline such that the user is able to define whether data are normalized to a common reference template, or alternatively, to age-specific fetal templates, see Serag *et al.* (2012). A linear

normalization is implemented via FLIRT (Jenkinson et al., 2002). After normalization, all volumes are spatially smoothed with a user-specified Gaussian kernel.

4.7 Flexible, Open Source Pipeline

Steps within this preprocessing pipeline have been merged into Python scripts utilizing Nipype (Gorgolewski et al., 2011). Nipype is a framework that aims to enable users to create more reliable and efficient data processing pipelines. For ease of transfer across datasets, the file organization required to run this pipeline conforms to the widely-used BIDS structure (Gorgolewski et al., 2016). Importantly, all software tools used within this preprocessing pipeline (AFNI, FSL, TensorFlow, Python, Nipype) are free and open-source. User determinations can be entered as to desired smoothing kernel size and normalization strategy, and an interface is provided for users to enter subjects' gestational ages. In addition, for users less familiar with Python, all commands can be implemented in a shell script, which can be run from the command line.

4.8 Quality Control

While this methodology employs fully automated techniques for preprocessing fetal resting-state fMRI data, manual quality assurance processes are advisable for key transition points throughout the pipeline. Specifically, our standard process includes initial review of raw time-series data in FSLeyes, screened as a movie. Criteria are that the brain is in the field of view and unobstructed by artifacts, and that within the time series, there are periods of minimal fetal movement. We exclude data not meeting these criteria. However, the majority of Wayne State University data passes this stage because long scan durations are used, and fetuses rapidly cycle through quiescent states.

Additional steps in the quality control protocol are implemented after auto-masking, realignment, and normalization (shown in Figure 5). At these stages timeseries data are again visually inspected using FSLeyes to assure that no errors were introduced during these stages of preprocessing. Additionally, the pipeline was constructed to generate quality control parameters, including realignment parameters, motion plots, and metrics from the fsl_motion_outliers command, that are useful for process evaluation and statistical modeling.

Acknowledgements

The authors thank Sahi Karra, Nedda Elewa, Tarek Bazzi, Tahir Khan, Nourhan Hamadi, Bryan Turman, Allison Li, Imran Sheikh, and Sophia Neuenfeldt for their time spent manually generating fetal brain masks, Pavan Jella for assistance in data acquisition, and Lauren Grove for the mentoring and advice while writing this manuscript. The authors also thank participant families who generously shared their time.

Author Contributions

Conceptualization: SR, MA; Methodology: SR, PS, MA, JH,; Formal Analysis: SR, PS, MA; Data Curation: SR, JH, DS; Writing – Original Draft: SR, PS, JH, MT, CS; Writing – Reviewing and Editing; SR, PS, MA, JH, MV, JW, DS, MT, CS; Visualization: SR, JH; Supervision: MV, JW, MT, CS; Funding Acquisition: MT.

References

Anderson, A. L., & Thomason, M. E. (2013). Functional plasticity before the cradle: A review of neural functional imaging in the human fetus. *Neuroscience & Biobehavioral Reviews*, *37*(9, Part B), 2220–2232. https://doi.org/10.1016/j.neubiorev.2013.03.013

Benkarim, O. M., Sanroma, G., Zimmer, V. A., Muñoz-Moreno, E., Hahner, N., Eixarch, E., Piella, G. (n.d.). Toward the automatic quantification of in utero brain development in 3D structural MRI: A review. *Human Brain Mapping*, *38*(5), 2772–2787. https://doi.org/10.1002/hbm.23536

Biegon, A., & Hoffmann, C. (2014). Quantitative magnetic resonance imaging of the fetal brain in utero: Methods and applications. *World Journal of Radiology*, *6*(8), 523–529. https://doi.org/10.4329/wjr.v6.i8.523

Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. *JAMA*.2017;318(6):517–518. doi:10.1001/jama.2017.7797

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *ArXiv:1511.06348 [Cs]*. Retrieved from http://arxiv.org/abs/1511.06348

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002). Image processing with neural networks—a review. *Pattern Recognition*, *35*(10), 2279–2301.

Fogtmann, M., Seshamani, S., Kroenke, C., Cheng, X., Chapman, T., Wilm, J., Studholme, C. (2014). A Unified Approach to Diffusion Direction Sensitive Slice Registration and 3-D DTI Reconstruction From Moving Fetal Brain Anatomy. *IEEE Transactions on Medical Imaging*, *33*(2), 272–289. https://doi.org/10.1109/TMI.2013.2284014

Garcia, K. E., Robinson, E. C., Alexopoulos, D., Dierker, D. L., Glasser, M. F., Coalson, T. S., … Bayly, P. V. (2018). Dynamic patterns of cortical expansion during folding of the preterm human brain. *Proceedings of the National Academy of Sciences*, *115*(12), 3156–3161. https://doi.org/10.1073/pnas.1715451115

Gholipour, A., Rollins, C. K., Velasco-Annis, C., Ouaalam, A., Akhondi-Asl, A., Afacan, O., Warfield, S. K. (2017). A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Scientific Reports*, *7*(1), 476. https://doi.org/10.1038/s41598-017-00525-w

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044. Retrieved from http://dx.doi.org/10.1038/sdata.2016.44

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*, 13. http://doi.org/10.3389/fninf.2011.00013

Hoo-Chang, S., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

Huang, H., & Vasung, L. (2014). Gaining Insight of Fetal Brain Development with Diffusion MRI and Histology. *International Journal of Developmental Neuroscience□: The Official Journal of the International Society for Developmental Neuroscience*, *32*, 11–22. http://doi.org/10.1016/j.ijdevneu.2013.06.005

Ison, M., Donner, R., Dittrich, E., Kasprian, G., Prayer, D., & Langs, G. (2012). Fully automated brain extraction and orientation in raw fetal MRI. In *Workshop on Paediatric and Perinatal Imaging, MICCAI* (pp. 17–24).

Jakab, A., Kasprian, G., Schwartz, E., Gruber, G. M., Mitter, C., Prayer, D., ... Langs, G. (2015). Disrupted developmental organization of the structural connectome in fetuses with corpus callosum agenesis. *NeuroImage*, *111*, 277–288. https://doi.org/10.1016/j.neuroimage.2015.02.038

Jakab, A., Schwartz, E., Kasprian, G., Gruber, G. M., Prayer, D., Schöpf, V., & Langs, G. (2014). Fetal functional imaging portrays heterogeneous development of emerging human brain networks. *Frontiers in Human Neuroscience*, *8*, 852. https://doi.org/10.3389/fnhum.2014.00852

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841.2002), pp. 825{841. issn : 1053-8119.
Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980 [Cs]*. Retrieved from http://arxiv.org/abs/1412.6980

Klinder, T., Wendland, H., Wachter-Stehle, I., Roundhill, D., & Lorenz, C. (2015). Adaptation of an articulated fetal skeleton model to three-dimensional fetal image data. In *Medical Imaging 2015: Image Processing* (Vol. 9413, p. 94130Q). International Society for Optics and Photonics. https://doi.org/10.1117/12.2081139

Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S. J., Doria, V., Serag, A., Rueckert, D. (2011). A dynamic 4D probabilistic atlas of the developing brain. *NeuroImage*, *54*(4), 2750–2763. https://doi.org/10.1016/j.neuroimage.2010.10.019

Link, D., Braginsky, M. B., Joskowicz, L., Ben Sira, L., Harel, S., Many, A., ... Ben Bashat, D. (2017). Automatic Measurement of Fetal Brain Development from Magnetic Resonance Imaging: New Reference Data. *Fetal Diagnosis and Therapy*. https://doi.org/10.1159/000475548

Makropoulos, A., Counsell, S. J., & Rueckert, D. (2017). A review on automatic fetal and neonatal brain MRI segmentation. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.06.074

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Rueckert, D. (2018). The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2018.01.054

Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., Lein, E. S. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, *508*(7495), 199–206. https://doi.org/10.1038/nature13185

Nunes, R. G., Ferrazzi, G., Price, A. N., Hutter, J., Gaspar, A. S., Rutherford, M. A., & Hajnal, J. V. (2018). Inner-volume echo volumar imaging (IVEVI) for robust fetal brain imaging. *Magnetic Resonance in Medicine*, *80*(1), 279–285. https://doi.org/10.1002/mrm.26998

Ouyang, A., Jeon, T., Sunkin, S. M., Pletikos, M., Sedmak, G., Sestan, N., ... Huang, H. (2015). Spatial Mapping of Structural and Connectional Imaging Data for the Developing Human Brain with Diffusion Tensor Imaging. *Methods (San Diego, Calif.)*, *0*, 27–37. https://doi.org/10.1016/j.ymeth.2014.10.025

Ouyang, M., Dubois, J., Yu, Q., Mukherjee, P., & Huang, H. (2018). Delineation of early brain development from fetuses to infants with diffusion MRI and beyond. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2018.04.017
pp. 1288{1292. doi : 10.1109/ISBI.2018.8363807 .

Qiu, A., Mori, S., & Miller, M. I. (2015). Diffusion Tensor Imaging for Understanding Brain Development in Early Life. *Annual Review of Psychology*, *66*(1), 853–876. https://doi.org/10.1146/annurev-psych-010814-015340

Rajchl, M., Lee, M. C. H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Rueckert, D. (2016). DeepCut: Object Segmentation from Bounding Box Annotations using Convolutional Neural Networks. *ArXiv:1605.07866 [Cs]*. Retrieved from http://arxiv.org/abs/1605.07866

Robinson, A. J., & Ederies, M. A. (2018). Fetal neuroimaging: an update on technical advances and clinical findings. *Pediatric Radiology*, *48*(4), 471–485. https://doi.org/10.1007/s00247-017-3965-z

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Auto-context Convolutional Neural Network (Auto-Net) for Brain Extraction in Magnetic Resonance Imaging. *IEEE Transactions on Medical Imaging*, *PP*(99), 1–1. https://doi.org/10.1109/TMI.2017.2721362

Scheinost, D. et al. "A fetal fMRI specic motion correction algorithm using 2nd order edge features". In:IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)

Schöpf, V., Kasprian, G., Brugger, P. C., & Prayer, D. (2012). Watching the fetal brain at 'rest.' *International Journal of Developmental Neuroscience*, *30*(1), 11–17. https://doi.org/https://doi.org/10.1016/j.ijdevneu.2011.10.006

Schuler, A.-L., Bartha-Doering, L., Jakab, A., Schwartz, E., Seidl, R., Kienast, P., Kasprian, G. (2018). Tracing the structural origins of atypical language representation: consequences of prenatal mirror-imaged brain asymmetries in a dizygotic twin couple. *Brain Structure and Function*, *223*(8), 3757–3767. https://doi.org/10.1007/s00429-018-1717-y

Serag, A., Aljabar, P., Ball, G., Counsell, S. J., Boardman, J. P., Rutherford, M. A., Rueckert, D. (2012). Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, *59*(3), 2255–2265. https://doi.org/10.1016/j.neuroimage.2011.09.062

Serag, A., Macnaught, G., Denison, F. C., Reynolds, R. M., Semple, S. I., & Boardman, J. P. (2017). Histograms of Oriented 3D Gradients for Fully Automated Fetal Brain Localization and Robust Motion Correction in 3□T Magnetic Resonance Images. *BioMed Research International*, *2017*. https://doi.org/10.1155/2017/3956363

Seshamani, S., Cheng, X., Fogtmann, M., Thomason, M. E., & Studholme, C. (2014). A Method for handling intensity inhomogenieties in fMRI sequences of moving anatomy of the early developing brain. *Medical Image Analysis*, *18*(2), 285–300. https://doi.org/10.1016/j.media.2013.10.011

Seshamani, S., Blazejewska, A. I., Gatenby, C., Mckown, S., Caucutt, J., Dighe, M., & Studholme, C. (2015). Comparing consistency of R2* and T2*-weighted BOLD analysis of resting state fetal fMRI. In *Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging* (Vol. 9417, p. 94170N). International Society for Optics and Photonics. https://doi.org/10.1117/12.2082236

Shattuck, D. W., & Leahy, R. M. (2002). BrainSuite: An automated cortical surface identification tool. *Medical Image Analysis*, 14.

Shi, F., Yap, P.-T., Fan, Y., Gilmore, J. H., Lin, W., & Shen, D. (2010). Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation. *NeuroImage*, *51*(2), 684–693.

Song, L., Mishra, V., Ouyang, M., Peng, Q., Slinger, M., Liu, S., & Huang, H. (2017). Human Fetal Brain Connectome: Structural Network Development from Middle Fetal Stage to Birth. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00561

Studholme, C. (2011). Mapping Fetal Brain Development In Utero Using Magnetic Resonance Imaging: The Big Bang of Brain Mapping. *Annual Review of Biomedical Engineering*, *13*(1), 345–368. https://doi.org/10.1146/annurev-bioeng-071910-124654

Studholme, C. (2015). Mapping the developing human brain in utero using quantitative MR imaging techniques. *Seminars in Perinatology*, *39*(2), 105–112. https://doi.org/10.1053/j.semperi.2015.01.003

Takahashi, E., Folkerth, R. D., Galaburda, A. M., & Grant, P. E. (2012). Emerging cerebral connectivity in the human fetal brain: an MR tractography study. *Cerebral Cortex (New York, N.Y.: 1991)*, *22*(2), 455–464. https://doi.org/10.1093/cercor/bhr126

Thomason, M. E. (2018). Structured Spontaneity: Building Circuits in the Human Prenatal Brain. *Trends in Neurosciences*, *41*(1), 1–3. https://doi.org/10.1016/j.tins.2017.11.004

Thomason, M. E., Brown, J. A., Dassanayake, M. T., Shastri, R., Marusak, H. A., Hernandez-Andrade, E., Romero, R. (2014). Intrinsic Functional Brain Architecture Derived from Graph Theoretical Analysis in the Human Fetus. *PLOS ONE*, *9*(5), e94423. https://doi.org/10.1371/journal.pone.0094423

Thomason, M. E., Dassanayake, M. T., Shen, S., Katkuri, Y., Alexis, M., Anderson, A. L., Romero, R. (2013). Cross-Hemispheric Functional Connectivity in the Human Fetal Brain. *Science Translational Medicine*, *5*(173), 173ra24-173ra24. https://doi.org/10.1126/scitranslmed.3004978

Thomason, M. E., Grove, L. E., Lozon, T. A., Vila, A. M., Ye, Y., Nye, M. J., Romero, R. (2015). Age-related increases in long-range connectivity in fetal functional neural connectivity networks in utero. *Developmental Cognitive Neuroscience*, *11*, 96–104. https://doi.org/10.1016/j.dcn.2014.09.001

Thomason, M. E., Scheinost, D., Manning, J. H., Grove, L. E., Hect, J., Marshall, N., Romero, R. (2017). Weak functional connectivity in the human fetal brain prior to preterm birth. *Scientific Reports*, *7*, 39286. https://doi.org/10.1038/srep39286

Tommasi, T., Patricia, N., Caputo, B., & Tuytelaars, T. (2015). A Deeper Look at Dataset Bias. *ArXiv:1505.01257 [Cs]*. Retrieved from http://arxiv.org/abs/1505.01257

Tourbier, S., Velasco-Annis, C., Taimouri, V., Hagmann, P., Meuli, R., Warfield, S. K., Gholipour, A. (2017). Automated template-based brain localization and extraction for fetal brain MRI reconstruction. *NeuroImage*, *155*, 460–472. https://doi.org/10.1016/j.neuroimage.2017.04.004

van den Heuvel, M. I., & Thomason, M. E. (2016). Functional Connectivity of the Human Brain in Utero. *Trends in Cognitive Sciences*, *20*(12), 931–939. https://doi.org/10.1016/j.tics.2016.10.001

van den Heuvel, M. I., Turk, E., Manning, J. H., Hect, J., Hernandez-Andrade, E., Hassan, S. S., ... Thomason, M. E. (2018). Hubs in the human fetal brain network. *Developmental Cognitive Neuroscience*, *30*, 108–115. https://doi.org/10.1016/j.dcn.2018.02.001

Vasung, L., Abaci Turk, E., Ferradal, S. L., Sutin, J., Stout, J. N., Ahtam, B., Grant, P. E. (2018). Exploring early human brain development with structural and physiological neuroimaging. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2018.07.041

Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a PhysicianHumanism and Artificial Intelligence. *JAMA*. 2018;319(1):19–20. doi:10.1001/jama.2017.19198

Wright, R., Kyriakopoulou, V., Ledig, C., Rutherford, M. A., Hajnal, J. V., Rueckert, D., & Aljabar, P. (2014). Automatic quantification of normal cortical folding patterns from fetal brain MRI. *NeuroImage*, *91*, 21–32. https://doi.org/10.1016/j.neuroimage.2014.01.034

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Confounding variables can degrade generalization performance of radiological deep learning models. *ArXiv:1807.00431 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1807.00431

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014* (pp. 818–833). Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53