Fetal Tissue Annotation Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Fetal Tissue Annotation Challenge

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

FeTA

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Congenital disorders are one of the leading causes of infant mortality worldwide [1]. In-utero MRI of the fetal brain has started to emerge as a valuable tool for investigating the neurological development of fetuses with congenital disorders. Fetal MRI can aid in the future development of clinical risk stratification tools for early interventions, treatments, and clinical counseling. Moreover, fetal MRI is a powerful tool to portray the complex neurodevelopmental events during human gestation, which remain to be completely characterized. Acquisition and analysis of in-utero MRI of the fetal brain requires collaboration from specialized clinical centers because image cohorts of this vulnerable patient populations are small and heterogeneous (e.g. variability in image acquisition parameters between sites). In the majority of specialized clinical centers working with fetal MRI, assessments are performed using only 2D biometric measurements derived from thick 2D slice acquisitions, although recent work has demonstrated the ability to perform these measurements in 3D super-resolution reconstructed volumes [2], [3].

Automated biometry, segmentation, and quantification of the highly complex and rapidly changing brain morphology prior to birth in MRI data would improve the diagnostic process, as manual annotations are both time consuming and subject to human error and inter-rater variability. It is clinically relevant to analyze information such as the shape or volume of the developing brain structures, as many congenital disorders cause subtle changes to these tissue compartments [4], [5], [6], [7]. Existing growth data is mainly based on normally developing brains [2], [8], [9], and growth data for many pathologies and congenital disorders is lacking. Thus, robust methods for automatic quantification of the developing human brain across different scanners and image acquisition protocols would be a first step in performing such an analysis.

From a technical standpoint, there are many challenges that an automatic segmentation method of the fetal brain would need to overcome. During prenatal development, the physiology of the human brain changes while its structure undergoes developmental reorganization. In addition, the quality of the images is often poor due to fetal and maternal movement and imaging artefacts [10], [11], while partial volume effects frequently lead to blurring of boundary between tissues. Finally, compared to the healthy controls, structures of an abnormal fetal brain often have a different morphology. This can make it challenging for an automatic method to recognize what the structures are. The field of fetal MRI has so far been understudied due to challenges in imaging and due to the lack of public, curated, and annotated ground truth data. Site and MRI scanner harmonization, paired with automated and robust methods for MRI analyses are needed in order to increase sample size for adequate power of these studies.

In our first FeTA edition (FeTA 2021), we used the first publicly available dataset of fetal brain MRI images [12] to encourage participating teams to develop automatic brain tissue segmentation methods [13]. Based on its success, we extended the challenge in FeTA 2022 (https://feta.grand-challenge.org) by studying the generalizability of segmentation algorithms across different sites, acquired from different image acquisition protocols and MRI scanners) (paper submitted at IEEE TMI, awaiting associate editor assignment and submitted to Arxiv awaiting approval). In FeTA 2024 we aim to improve and extend the established FeTA Challenge in two ways: Firstly, we introduce a new task to automatically derive clinically relevant biometry typically used in practice for fetal evaluation. Secondly, following the recent rise in popularity of low-cost low-field MRI systems [14] which aim at democratizing MRI access world-wide, we extend the generalizability assessment of segmentation methods by including low-field (0.55T) MRI data. FeTA 2024 challenge is an important step towards the development of effective, domain-generalizable and reproducible methods for analyzing high resolution reconstructed MR images of the developing fetal brain from gestational week 21-36. It will include a new clinically relevant task on automated biometry measurements and data from five different sites and magnetic fields including recent low-field systems. Such new algorithms will have the potential to support clinicians in the assessment of fetal brain MRI in-utero in many ways: to better understand the underlying causes of congenital disorders and ultimately to guide the development of antenatal/postnatal guidelines and clinical risk stratification tools for early interventions, treatments, and care management decisions across hospitals and research institutions worldwide.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Fetal Brain MRI, Biometry, Segmentation, Multi-class, Domain Generalization, Congenital Disorders, Low Field MRI, Regression

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

Our challenge would be integrated as Thematic event within PIPPI 2024.

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect 15-30 submissions. Together with invited speakers, organizers this will make about 30-50 participants. Previous challenge editions in 2021 and 2022 gathered 21 and 17 submissions respectively, with around 35 to 50 participants each year.

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

We plan to coordinate a publication of the challenge results after the challenge, targeting publication at high impact factor journal (eg. Medical Image Analysis (FeTA 2021: https://doi.org/10.1016/j.media.2023.102833), IEEE Transactions on Medical Imaging (FeTA 2022 challenge paper currently submitted to Arxiv and IEEE TMI), or NeuroImage). All teams who submitted before the deadline and presented their results at MICCAI 2024 will be included in the paper. Each team is allowed three co-authors in this paper.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

This challenge will be on-site (having a hybrid setup (i.e. Zoom) will be considered if needed). We would need a seminar room with projector and screen, and a microphone on the day of the challenge.

TASK 1: Generalizable Fetal Brain Segmentation

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Congenital disorders are one of the leading causes of infant mortality worldwide [1]. In-utero MRI of the fetal brain has started to emerge as a valuable tool for investigating the neurological development of fetuses with congenital disorders. Fetal MRI can aid in the future development of clinical risk stratification tools for early interventions, treatments, and clinical counseling. Moreover, fetal MRI is a powerful tool to portray the complex neurodevelopmental events during human gestation, which remain to be completely characterized. Acquisition and analysis of in-utero MRI of the fetal brain requires collaboration from specialized clinical centers because image cohorts of this vulnerable patient populations are small and heterogeneous (e.g. variability in image acquisition parameters between sites). In the majority of specialized clinical centers working with fetal MRI, assessments are performed using only 2D biometric measurements derived from thick 2D slice acquisitions, although recent work has demonstrated the ability to perform these measurements in 3D super-resolution reconstructed volumes [2], [3].

Automated biometry, segmentation, and quantification of the highly complex and rapidly changing brain morphology prior to birth in MRI data would improve the diagnostic process, as manual annotations are both time consuming and subject to human error and inter-rater variability. It is clinically relevant to analyze information such as the shape or volume of the developing brain structures, as many congenital disorders cause subtle changes to these tissue compartments [4], [5], [6], [7]. Existing growth data is mainly based on normally developing brains [2], [8], [9], and growth data for many pathologies and congenital disorders is lacking. Thus, robust methods for automatic quantification of the developing human brain across different scanners and image acquisition protocols would be a first step in performing such an analysis.

From a technical standpoint, there are many challenges that an automatic segmentation method of the fetal brain would need to overcome. During prenatal development, the physiology of the human brain changes while its structure undergoes developmental reorganization. In addition, the quality of the images is often poor due to fetal and maternal movement and imaging artefacts [10], [11], while partial volume effects frequently lead to blurring of boundary between tissues. Finally, compared to the healthy controls, structures of an abnormal fetal brain often have a different morphology. This can make it challenging for an automatic method to recognize what the structures are. The field of fetal MRI has so far been understudied due to challenges in imaging and due to the lack of public, curated, and annotated ground truth data. Site and MRI scanner harmonization, paired with automated and robust methods for MRI analyses are needed in order to increase sample size for adequate power of these studies.

In our first FeTA edition (FeTA 2021), we used the first publicly available dataset of fetal brain MRI images [12] to encourage participating teams to develop automatic brain tissue segmentation methods [13]. Based on its success, we extended the challenge in FeTA 2022 (https://feta.grand-challenge.org) by studying the generalizability of segmentation algorithms across different sites, acquired from different image acquisition protocols and MRI scanners) (paper under review at IEEE TMI https://arxiv.org/abs/2402.09463). In FeTA 2024 we aim to improve and extend the established FeTA Challenge in two ways: Firstly, we introduce a new task to

automatically derive clinically relevant biometry typically used in practice for fetal evaluation. Secondly, following the recent rise in popularity of low-cost low-field MRI systems [14] which aim at democratizing MRI access world-wide, we extend the generalizability assessment of segmentation methods by including low-field (0.55T) MRI data. FeTA 2024 challenge is an important step towards the development of effective, domain-generalizable and reproducible methods for analyzing high resolution reconstructed MR images of the developing fetal brain from gestational week 21-36. It will include a new clinically relevant task on automated biometry measurements and data from five different sites and magnetic fields including recent low-field systems. Such new algorithms will have the potential to support clinicians in the assessment of fetal brain MRI in-utero in many ways: to better understand the underlying causes of congenital disorders and ultimately to guide the development of antenatal/postnatal guidelines and clinical risk stratification tools for early interventions, treatments, and care management decisions across hospitals and research institutions worldwide.

Keywords

List the primary keywords that characterize the task.

MRI, Fetal Brain, Tissue Segmentation, Multi-class, Domain Generalization, Congenital Disorders, Low Field MRI

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

The same as for the challenge.

b) Provide information on the primary contact person.

Meritxell Bach Cuadra (meritxell.bachcuadra@unil.ch)

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Previous challenges were held at MICCAI 2021 and MICCAI 2022. There were 21 and 17 submissions (from 20 and 16 teams) respectively. FeTA challenge at MICCAI 2024 would consolidate and further grow the research community surrounding it, and to broaden tasks. Specifically, we will incorporate a new task to automatically predict clinically relevant information of brain biometry and add a new low-field MRI dataset for evaluation.

Training data will be made available during the whole time, and after the challenge. The test data used during the testing phase of the challenge will remain private and completely hidden to challenge participants.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

N/A

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

feta.grand-challenge.org

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

The participants can use publicly available data as they wish but should document whatever is used in the description of their algorithm. Participants may modify the provided training data as they wish. This modification includes the generation of additional data by image synthesis or various data augmentation strategies (for example, using FABIAN [17]) as long as everything is documented, and synthetic data should be able to be made available upon request.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Members of the organizers institutes may participate in the challenge but are not eligible for awards.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

The top three ranking methods will be publicly named and awarded certificates and a small gift, such as chocolates (prize of previous editions was custom-made FeTA chocolate bars).

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The results will be announced publicly at the MICCAI 2024 challenge session, and will be posted on the challenge website. The teams with the top 10 algorithms will be informed earlier so they can prepare a presentation for the challenge session. The top 3-5 teams will be asked to prepare a 7-10 minute presentation, and the remaining 5-7 teams will be asked to prepare a short speed round presentation of 2 minutes.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)

- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Three authors per team who contributed to the design of the algorithm will be named co-author in the final challenge paper. Every participant can publish their algorithms and results independently after the challenge, but we request they cite the summary paper, and the data publication paper. The results and our corresponding evaluation of all participating teams will be made publicly available on the challenge website after the conference session.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants will create a Docker container with their algorithm, and provide this to the challenge organizers. Once the website is up, it will contain instructions on how to containerize the algorithms, and the organizers will provide support to the participants when requested. The organizers will run the Docker container on the test data set using publicly available evaluation code. The organizers will inform the participants if the Docker container fails to run, and allow the participants to provide a fix.

https://feta.grand-challenge.org/Submission/ (available after challenge registration)

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

No multiple submission will be allowed. The evaluation will be performed on the submitted Docker containers at the organizers institute. Resubmissions are only allowed in cases of technical errors with the Docker.
After the MICCAI challenge session, new submissions and updates may be made.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Training Data Release: May, 2024
Registration: Will be announced once the challenge is accepted
Docker Submission Deadline: 2 months before the challenge session
Top 10 teams informed: 2 weeks before the challenge, so they can prepare a presentation

Complete results will be announced at the challenge session

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Kispi Data: Mothers of the foetuses gave general or specific research consent for the further use of their data. The ethical committee of the Canton of Zurich, Switzerland approved the studies that collected and analysed the MRI data (Decision numbers: 2017 00885, 2016 01019, 2017 00167), and a waiver for an ethical approval was acquired for the release of a fully anonymous dataset.

CHUV Data: Mothers of all other foetuses included in the current work were scanned as part of their routine clinical. Data was retrospectively collected from acquisitions done between January 2013 to April 2021 from the MRI foetal brain database of our institution. All images were anonymised. This dataset is part of a larger research protocol approved by the ethics committee of the Canton de Vaud (decision number CERVD 2021 00124) for re-use of their data for research purposes and approval for the release of an anonymous dataset for non-medical reproducible research and open science purposes.

Vienna Data: The data has been acquired in the course of a retrospective single-center study and has been anonymised and approved by the ethics review board and data clearing department at the Medical University of Vienna, responsible for validating data privacy and sharing regulation compliance.

UCSF Data: Fetal MRI was clinically acquired and approved for anonymised retrospective studies for internal use studies (IRB 16 20619). A data sharing agreement and new IRB has bee set to expand the ethical approval for sharing anonymised data outside of the institution.

KCL Data: Fetal MRI was acquired at Kings College London and approved for sharing with interested academic researchers around the world by the Ethics Committee Dulwich (Ethics code 19 LO 0852). The data has been acquired during a prospective single-center study and has been fully anonymised in line with local procedures.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)

- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Custom Kispi data license (current FeTA license can be found here under Terms of Use: https://www.synapse.org/(dash)!Synapse:syn25649159/wiki/610007); non-commercial use only for the challenge purpose.
Custom Vienna Data Transfer Agreement that each participant must sign prior to receiving the data.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Evaluation code will be available on the challenge website.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

We will encourage participating teams to make their code/submission public. At the end of the challenge a Docker Hub page will be created to store the containers submitted by participants as it was done for the previous edition of FETA2022 (https://hub.docker.com/u/fetachallenge22).

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No sponsoring/funding is planned.
Only the organizers will have access to the test labels.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research, Diagnosis, Decision support

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort would be pregnant mothers who, after a screening ultrasound, have been clinically referred to a fetal MRI.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

In the challenge, a clinically acquired dataset will be used representing the target cohort. Fetal MRI brain scans that were acquired clinically, and reconstructed using a super-resolution method.
There are two subgroups in the dataset:
1. Fetuses with normal neurodevelopment
2. Fetuses with pathological neurodevelopment

We will include datasets from at least five different centers.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Magnetic Resonance Imaging: Several T2-weighted Fast Spin Echo (referred as ssFSE)/ Half fourier single-shot Turbo spin echo (HASTE) depending on the vendor) images were acquired for each subject in all three planes (with at least one image in each of the axial, sagittal, coronal planes) with resolution 0.5mm x 0.5mm x 3mm (Kispi, UCSF, Vienna) or 1.125mm x 1.125mm x 3mm (CHUV), or 1.5mm x 1.5mm x 4.5mm (KCL) and were combined together to reconstruct into a single high resolution volume of 0.5mm x 0.5mm x 0.5mm (or 1.125mm x 1.125mm x 1.125mm, or 1.2mm x 1.2mm x 1.2mm).

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Voxel-level segmentations of the fetal brain:
0: Background
1: External Cerebrospinal Fluid
2: Grey Matter
3: White Matter
4: Ventricles
5: Cerebellum
6: Deep Grey Matter
7: Brainstem

b) ... to the patient in general (e.g. sex, medical history).

Gestational age (GA) in weeks
Pathological or Neurotypical brain

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Fetal MRI Brain Scan.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Individual structures within the fetal brain (external cerebrospinal fluid, grey matter, white matter, ventricles, brain stem, cerebellum, deep grey matter)

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.

- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy, Precision, Generalisability

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Kispi Data: 1.5T and 3T clinical GE whole-body scanner (Signa Discovery MR450 and MR750) were used to acquire the data, either using an 8-channel cardiac coil or body coil without the use of maternal or fetal sedation.
CHUV: 1.5 T (MAGNETOM Aera, Siemens Healthcare, Erlangen, Germany), without the use of maternal or fetal sedation. Acquisitions were performed with an 18-channel body coil and a 32-channel spine coil.
Vienna: 1.5 T (Philips Ingenia/Intera, Best, the Netherlands) and 3 T magnets (Philips Achieva, Best, the Netherlands), without the use of maternal or fetal sedation. All acquisitions were performed using a cardiac coil.
UCSF: 3T GE Discovery MR750 or MR750W (wide bore) without the use of maternal or fetal sedation. Acquisitions were performed using a 32 channel GE cardiac coil.
KCL: 0.55T SIEMENS MAGNETOM FREE.MAX, Siemens Healthineers, Erlangen, Germany) without the use of maternal or fetal sedation. All acquisitions were performed with the contour L coil and the integrated spine coil in maternal supine position.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Kispi: T2-weighted SSFSE sequences were acquired with an in-plane resolution of 0.5mmx0.5mm and a slice thickness of 3 to 5mm. The sequence parameters were the following: TR: 2000-3500 ms, TE: 120 ms (minimum), flip angle: 90 degrees, sampling percentage 55%. Field of view (200-240 mm) and image matrix (1.5T: 256x224; 3T: 320x224) were adjusted depending on the gestational age (GA) and size of the fetus. Imaging plane was oriented relative to the fetal brain and axial, coronal and sagittal images were acquired.

CHUV: T2-weighted (T2W) Half-Fourier Acquisition Single-shot Turbo spin Echo (HASTE) sequences in the three orthogonal orientations; usually at least two acquisitions were performed in each orientation., TR/TE, 1200ms/90ms; flip angle, 6/23 90 degrees; echo train length, 224; echo spacing, 4.08 ms; field-of-view, 360 × 360mm2 ; voxel size, 1.13 × 1.13 × 3.00mm3 ; inter-slice gap, 10%, acquisition time between 26 to 36 seconds.

Vienna: For each case, at least 3 T2-weighted single-shot, fast spin echo (ssFSE) sequences (TE=80-140ms) in 3 orthogonal (axial, coronal, sagital) planes with reference to the fetal brain stem axis and/or the axis of the corpus callosum have been acquired using a 1.5 Tesla Philips Intera MR scanner. Overall, slice thickness was between 3mm and 5mm (gap 0.3-1mm), pixel spacing 0.65-1.17mm, acquisition time between 13.46 and 41.19 seconds.

UCSF: At least 3 T2-weighted ssFSE sequences were acquired with one scan per orientation (sagittal, axial, coronal) with the following parameters: 240 mm FOV with 512x512 matrix gives in plane resolution of around 0.5x0.5 mm with 3 mm slice thickness. TR is 2000-3500 ms, TE > 100 ms, 90 deg flip angle.

KCL: Half-Fourier Acquisition Single-shot Turbo spin Echo (HASTE) sequences acquired in the three orthogonal orientations; usually the brain is covered in all three radiological presentations with three additional whole uterus stacks performed with a higher number of slices to encompass whole uterus in the FoV; acquisition parameters were TR/TE, 2500ms/106ms; flip angle, 180 deg; field-of-view, 450 × 450mm2 with a 304x304 pixels base resolution; voxel size, 1.5 × 1.5 × 4.5mm3 ; acquisition time between 64 to 122 seconds.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Kispi: The data was acquired at the University Childrens Hospital Zurich in Zurich, Switzerland.
CHUV: The data was acquired at the Lausanne University Hospital in Lausanne, Switzerland.
Vienna: The data was acquired at the General Hospital Vienna/Medical University of Vienna in Vienna, Austria
UCSF: The data was acquired at the University of California San Francisco
KCL: The data was acquired at St Thomas Hospital in London, United Kingdom.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Trained radiographers acquired the data using clinically defined protocols.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Each case consists of a 3D super-resolution reconstruction of a fetal brain (256x256x256 voxels). Training cases have an annotated label map corresponding to 7 different brain tissue types and 5 biometry measurements. Each case will have a corresponding gestational age in weeks, as well as a label if it is a neurotypical fetal brain or a pathological fetal brain.
Neither label maps nor biometry measurements are provided for the test cases.

b) State the total number of training, validation and test cases.

Kispi:
Training/Validation cases: 80 3D volumes
Test cases: 40 3D volumes

Vienna:

Training/Validation cases: 40 3D volumes
Test cases: 40 3D volumes

CHUV:
Test cases: 40 3D volumes

UCSF:
Test cases 40 3D volumes

KCL:
Test cases 20 3D volumes

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

At the moment, the Kispi, Vienna training and testing datasets and UCSF and CHUV testing datasets have been prepared, and were the datasets used in FeTA 2022. We aim to expand the dataset for this new challenge, by adding 20 additional cases at low-field MRI (already acquired) and biometry annotations for all the training and testing subjects (N=300), as this is what we can commit to providing by the time the challenge goes live. By incorporating extra cases obtained from a field strength different from those in the training set, we can assess how well the submitted algorithms handle changes in image properties resulting from reconstructing images from low-field MR datastacks. This evaluation helps us gauge the algorithms' robustness to domain shifts for both segmentation and biometry estimation paving the way to evaluating clinical applicability of the low-field MRI, comparing it in our unique dataset to two other higher field strengths.

Biometric measurements are good markers of fetal brain maturation and growth and are a fundamental basis for the clinical diagnosis of developmental and acquired brain abnormalities [18]. Quantifying brain development in relation to reference charts, using a biometric measurement, is a routine procedure in prenatal diagnosis using MRI. Yet, fetal brain MRI biometric analysis is commonly performed on low quality images with low confidence and uncertainty. Previous studies explored the biometry between low-resolution clinical series and SR reconstruction. It was shown that super-resolution reconstruction enables similar fetal brain biometrics on healthy patients, while improving the confidence level of the measurements [2], [3], [19].

Existing studies have shown that case numbers such as what we are providing are sufficient to train U-Nets to segment the fetal brain from maternal tissue, as well as the fetal brain into different tissue classes [12], [20], [21]. In addition, fetal MRI case numbers per hospital can be fairly small, making it a challenge for many hospitals to create such a uniform in-house single-center dataset. There are only a few centers worldwide where fetal MRI case numbers are in the range of a few hundred per year. For comparison, in our hospital, 11.9 fetal MRI scans have been done per month over the past two years, while other university hospitals in Switzerland reported performing on average 2-3 fetal MRI scans, including neuro sequences per month. This renders the collection of data challenging and makes our data unique.

Our dataset is the first publicly released fetal brain dataset consisting of manually annotated volumes with multiple classes, making this a unique dataset.

Finally, the distribution of GAs and number of pathologic subjects is equal between the test and training data.

In addition, we are increasing the number of sites from four to five, where we provide two sites in the training dataset so that the participants are able to train and test their algorithms on two sites, while keeping three sites held back to be part of the testing dataset to determine how well each algorithm performs on an unknown dataset.

We are not providing an explicit validation set, it is up to each team to decide on their own based on the data given.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

In the training and test set, fetal brains with a variety of pathologies of varying severities will be included (such as spina bifida and ventriculomegaly), as well as fetal brains with normal neurodevelopment. There will be more pathological fetal brains included than normal, as in a clinical setting it is more common to see pathological brains. In addition, the main goal is assessment of brain development through the segmentation and biometry of fetal brains with congenital disorders for further analysis. In both of the training and test sets, a range of gestational ages (20-36 weeks) is included. The split/distribution of pathologies and gestational ages will be equal between the training and test sets. The fetal brain undergoes a large variety of changes throughout gestation such as brain volume increase, gyrification, neuronal migration, and synaptogenesis. As a result, the tissue contrast, especially between the grey matter and white matter is changing throughout all gestation, adding more complexity to the segmentation. The tissues become either more distinguishable (such as between the deep grey matter and white matter) or less (such as between the white matter and grey matter) throughout gestation. Therefore, we aim to include as equal case numbers as possible for each gestational week.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Each label was created separately, with the annotator segmenting every second to every third slice for each label in the axial plane, except for the cerebellum and the brainstem/spinal cord, which were segmented in the sagittal plane. The final label map was created by post-processing these sparse annotations to create a single fully segmented fetal brain for each subject. For interpolating the sparsely annotated label maps, we used the Python implementation of the ITK nD Morphological Contour Interpolation algorithm, enforcing interpolation along the plane the given structure was annotated. After post-processing, each fetal brain was reviewed by an expert and small corrections were made either on the original annotations or the reconstructed full fetal brain annotation (see section Instructions to the annotators for more details).

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The annotation protocol can be found within the dataset in the documentation folder (FetalAnnotationGuideline.pdf):
http://neuroimaging.ch/sites/default/files/FetalAnnotationGuideline.pdf

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Individuals with experience in segmenting medical images and performing biometric measurements completed this task. The annotators were either radiographers with experience in MRI segmentation, or resident physicians with 5 to 10 years of experience in MRI.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

For each subject, we manually reviewed the acquired fetal brain images for quality in order to compile a stack of images. Each stack consisted of at least one brain scan in each orientation, with more scans included when available. The number of scans in each stack ranged between 3 and 13. Every image in the stack was then reoriented to a standard plane and a mask was created of the fetal brain using a semi-automated atlas-based custom MeVisLab (MeVis Medical Solutions AG, Bremen, Germany) module. SR reconstruction [23]-[27] was then performed based on each subjects stack of images and brain masks, creating a 3D SR volume of brain morphology with an isotropic resolution of 0.5mm*0.5mm*0.5mm (Kispi) or 1.125mm*1.125mm*1.125mm (CHUV). Each image was histogram-matched using Slicer [28], and zero-padded to be 256x256x256 voxels.

In Vienna, an alternative pipeline was followed [29]. In each scanning session between 3 and 7 acquisitions of a fetus were obtained in 3 orthogonal directions. The preprocessing pipeline consists of a data denoising step [30], followed by an in-plane super resolution [31] and automatic brain masking step [26] and concludes with a single 0.5 mm isotropic slice-wise motion correction and volumetric super-resolution reconstruction. Subsequently, the resulting volumes are rigidly aligned to a common reference space [32].

UCSF subject scans were manually reviewed, and the good quality stacks were chosen for super resolution reconstruction, resulting in 0.8 mm3 isotropic 3D SR volume of brain morphology [26]. Each reconstruction was zero-padded to 256x256x256 and reoriented to a standard viewing plane.

For the KCL data, all stacks (between 6 and 9 available) were manually reviewed for quality. These with sufficient quality were automatically masked and then reconstructed to 1.2 mm isotropic resolution 3D volumes using SVRTK [25] and aligned to standard reference space.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Sources of error in the image annotation depend mostly on the quality of the super-resolution reconstruction of the fetal brain. However, even the best quality reconstructions may contain annotator error. Errors can originate from:
- poor judgement of anatomical borders between certain developing brain structures, such as the cortical plate and subcortical white matter
- suboptimal image quality after reconstruction
- Annotations were made mainly in the axial plane, leading to some noisy labels when looking at the coronal and sagittal planes

An investigation on the inter-rater variability was performed with 3 annotators annotating 9 volumes, with high and medium quality reconstructions having a Dice score of 0.84 +/- 0.09 between the three raters, and low-quality reconstructions having a Dice score of 0.53+/- 0.24 between raters, averaged across all labels.

b) In an analogous manner, describe and quantify other relevant sources of error.

None.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Four complementary evaluation metrics will be used to compute the rankings:
- Dice similarity coefficient (DSC) to measure the spatial overlap [31] [32]
- the 95th percentile of Hausdorff distance (HD95) (https://github.com/deepmind/surface-distance) to quantify the contour distance.
- Volume similarity to measure the volume difference [31]
- the Euler characteristic (EC), defined k-dimensional Betti numbers to quantify topological correctness. ([33][34]).

In addition, intracranial volume will be calculated, but not used in the rankings. Intracranial volume will be determined by adding all labels together.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

As the task is a segmentation task, the DSC was chosen, as it is the most popular segmentation metric. However, we would like not just an overlap metric, but we are also interested in the shape and volume, therefore we will include the HD95 (shape) and VS (volume), and the final ranking will take all three metrics into account. We will include a topological metric as well, given the preliminary analysis we conducted on topology in our previous challenge (FeTA 2022, https://arxiv.org/abs/2402.09463).

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

This ranking system was developed in order to take four different metric types equally into account. We also wanted to determine not just average ranking, but see if algorithms performed better in when the image was high quality vs low quality, as well as how well they perform on the pathological vs the neurotypical brains.

b) Describe the method(s) used to manage submissions with missing results on test cases.

The only possibility for missing data would be if an algorithm does not find any of one label in the final label map, or if the entire label map is empty, or if biometry segments are missing or equal 0.

If there are missing results, the worst possible value will be used. For example, if a label does not exist in the label map, it will receive a DSC and VS of 0, and the HD95 will be double the max value of the other algorithms submitted (to ensure it is ranked last for that sub-ranking).

c) Justify why the described ranking scheme(s) was/were used.

The two tasks will have independent rankings.
For fetal tissue segmentation task, a final score incorporating the four metrics (DSC, HD95, VS, EC) will be determined. All metrics will be calculated for each label within each of the corresponding predicted label maps of the fetal brain volumes in the testing set. The mean and standard deviation of each label will be calculated, and the participating algorithms will be ranked from low to high (HD95, EC), where the lowest score receives the highest scoring rank (best), and from high to low (DSC, VS), where the highest value will receive highest scoring rank (best). For each label, the four rankings were added together, and the algorithm with the highest ranking was ranked first.

Results of the challenge will be run through the ChallengeR toolkit, specifically designed to calculate and display imaging challenge results [38].

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The description of missing data handling can be found above.
The mean and standard deviation of each method will be calculated using both an average of all labels as well as individually, using the ranking method as described above.

b) Justify why the described statistical method(s) was/were used.

This ranking system was developed in order to take four different metric types equally into account. We also wanted to determine not just average ranking, but see if algorithms performed better in when the image was high quality vs low quality, as well as how well they perform on the pathological vs the neurotypical brains.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Inter-algorithm variability may be analyzed in the final paper, as well as an in-depth analysis as to why some algorithms performed better than others (potential problems/biases that may be present).

Algorithms will be evaluated in the categories Non Pathological cases and Pathological cases, as well as in relation to the SR image quality, with the identical ranking scheme for each category. For further analysis, we will also manually rate the quality of each image according to the following criteria: geometrical artifacts (e.g. stripes), topological artefacts (e.g. discontinuous cortical ribbon, etc.), blurring, noise, bias field and contrast between tissues, as well as a global subjective quality rating [10], [11].

In addition, we plan to analyse the performance of the algorithms based on gestational age. The structure of the foetal brain changes greatly throughout development, especially in the cortex where there is increased cortical complexity, cortical specification (blurring of white matter and grey matter border) and partial volumes (blurring of white matter/grey matter border because of narrow gyri). To do so we will explore one additional metric (not included in the final ranking) for the segmentation task. Normalized Dice score (nDSC) is an adaptation of DSC which scales the precision at a fixed recall rate to tackle the bias of occurrence rate of the positive class in the ground-truth [35]. Since in our segmentation task we are dealing with multiple structures of different expected volumes as well as subjects with different gestation ages (that also affect target structure volumes) this metric provides an unbiased estimation of the volumetric segmentation quality independent of the target volumes distribution. We plan also on comparing the final rankings using the HD95 as well as the ASSD in order to see if there is an impact based on the surface metric used.
We plan to split our cohort into three age groups, and will analyse the submitted algorithms within these two groups to see if gestational age impacts the success of a segmentation algorithm or systematic biometry estimation errors.

In summary, we plan on analysing the same subgroups of the age separated cohorts using the same metrics as in the main ranking, as well as using the additional categories Non Pathological cases, Pathological cases and regarding the estimated SR image quality. We will also explore the nDSC metric as function of GA, and the relation of 3D volumetry with 2D biometry. All this analysis will be separate from the main ranking.

TASK 2: Fetal Brain Biometry

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Same as Task1.

Keywords

List the primary keywords that characterize the task.

MRI, Fetal Brain, Biometry, Regression, Domain Generalization, Congenital Disorders, Low Field MRI

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

The same as for the challenge.

b) Provide information on the primary contact person.

Meritxell Bach Cuadra (meritxell.bachcuadra@unil.ch)

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

The same as above.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

N/A

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

feta.grand-challenge.org

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

The participants can use publicly available data as they wish but should document whatever is used in the description of their algorithm. Participants may modify the provided training data as they wish. This modification includes the generation of additional data by image synthesis or various data augmentation strategies (for example, using FABIAN [17]) as long as everything is documented, and synthetic data should be able to be made available upon request.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Members of the organizers institutes may participate in the challenge but are not eligible for awards.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

The top three ranking methods will be publicly named and awarded certificates and a small gift, such as chocolates (prize of previous editions was custom-made FeTA chocolate bars).

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The results will be announced publicly at the MICCAI 2024 challenge session, and will be posted on the challenge website. The teams with the top 10 algorithms will be informed earlier so they can prepare a presentation for the challenge session. The top 3-5 teams will be asked to prepare a 7-10 minute presentation, and the remaining 5-7 teams will be asked to prepare a short speed round presentation of 2 minutes.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Three authors per team who contributed to the design of the algorithm will be named co-author in the final challenge paper. Every participant can publish their algorithms and results independently after the challenge, but we request they cite the summary paper, and the data publication paper. The results and our corresponding evaluation of all participating teams will be made publicly available on the challenge website after the conference session.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants will create a Docker container with their algorithm, and provide this to the challenge organizers. Once the website is up, it will contain instructions on how to containerize the algorithms, and the organizers will provide support to the participants when requested. The organizers will run the Docker container on the test data set using publicly available evaluation code. The organizers will inform the participants if the Docker container fails to run, and allow the participants to provide a fix.

https://feta.grand-challenge.org/Submission/ (available after challenge registration)

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

No multiple submission will be allowed. The evaluation will be performed on the submitted Docker containers at the organizers institute. Resubmissions are only allowed in cases of technical errors with the Docker.
After the MICCAI challenge session, new submissions and updates may be made.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Idem as Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Idem as Task 1.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

The same as above, non-commercial use only for the challenge purpose (to be further confirmed).

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Evaluation code will be available on the challenge website.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

We will encourage participating teams to make their code/submission public. At the end of the challenge a Docker Hub page will be created to store the containers submitted by participants as it was done for the previous edition of FETA2022 (https://hub.docker.com/u/fetachallenge22).

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No sponsoring/funding is planned.
Only the organizers will have access to the test labels.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis

- Research
- Screening
- Training
- Cross-phase
-

Research, Diagnosis, Decision support, Screening, Education

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Regression

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort would be pregnant mothers who, after a screening ultrasound, have been clinically referred to a fetal MRI.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Idem as Task 1.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Idem as Task 1.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Biometric measures of the fetal brain:
length of the corpus callosum (LCC), height of the vermis (HV), brain biparietal diameter (bBIP_ax), skull biparietal diameter (sBIP_ax), maximum transverse cerebellar diameter (TCD_cor)

b) ... to the patient in general (e.g. sex, medical history).

Gestational age (GA) in weeks
Pathological or Neurotypical brain

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Fetal MRI Brain Scan.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Biometric measures within the fetal brain (Length of corpuscollosum (LCC), height of the vermis (HV), brain biparietal diameter (bBIP_ax), skull biparietal diameter (sBIP_ax), transverse cerebellar diameter (TCD_cor)).

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy, Precision, Generalisability

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Idem as Task 1.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Idem as Task 1.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Idem as Task 1.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Trained radiographers acquired the data using clinically defined protocols.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Each case consists of a 3D super-resolution reconstruction of a fetal brain (256x256x256 voxels). Training cases have an annotated label map corresponding to 7 different brain tissue types and 5 biometry measurements. Each case will have a corresponding gestational age in weeks, as well as a label if it is a neurotypical fetal brain or a pathological fetal brain.
Neither label maps nor biometry measurements are provided for the test cases.

b) State the total number of training, validation and test cases.

Idem as Task 1.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Idem as above.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Idem as Task 1.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Quantitative biometry 2D evaluation is performed manually by an experienced neuroradiologist and following published guidelines ([2], [17], [20]). Length of the corpus callosum (LCC) and height of the vermis (HV) are measured in the mid-sagittal plane. Brain biparietal diameter (bBIP_ax), defined as the maximal brain diameter, and skull biparietal diameter (sBIP_ax), defined as the inner-to-inner table maximal skull diameter, are both measured in the transverse plane through the atria. Maximum transverse cerebellar diameter (TCD_cor) is measured in the coronal plane.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

https://unils-my.sharepoint.com/:b:/g/personal/meritxell_bachcuadra_unil_ch/ERBHlllq6lJFguOyxB299r0B5OIGfoj dqMLIMKzEScQowg?e=DwEkNa

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Individuals with experience in segmenting medical images and performing biometric measurements completed this task. The annotators were either radiographers with experience in MRI segmentation, or resident physicians with 5 to 10 years of experience in MRI.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

For each subject, we manually reviewed the acquired fetal brain images for quality in order to compile a stack of images. Each stack consisted of at least one brain scan in each orientation, with more scans included when available. The number of scans in each stack ranged between 3 and 13. Every image in the stack was then reoriented to a standard plane and a mask was created of the fetal brain using a semi-automated atlas-based custom MeVisLab (MeVis Medical Solutions AG, Bremen, Germany) module. SR reconstruction [23]-[27] was then performed based on each subjects stack of images and brain masks, creating a 3D SR volume of brain morphology with an isotropic resolution of 0.5mm*0.5mm*0.5mm (Kispi) or 1.125mm*1.125mm*1.125mm (CHUV). Each image was histogram-matched using Slicer [28], and zero-padded to be 256x256x256 voxels.

In Vienna, an alternative pipeline was followed [29]. In each scanning session between 3 and 7 acquisitions of a fetus were obtained in 3 orthogonal directions. The preprocessing pipeline consists of a data denoising step [30], followed by an in-plane super resolution [31] and automatic brain masking step [26] and concludes with a single 0.5 mm isotropic slice-wise motion correction and volumetric super-resolution reconstruction. Subsequently, the resulting volumes are rigidly aligned to a common reference space [32].

UCSF subject scans were manually reviewed, and the good quality stacks were chosen for super resolution reconstruction, resulting in 0.8 mm3 isotropic 3D SR volume of brain morphology [26]. Each reconstruction was zero-padded to 256x256x256 and reoriented to a standard viewing plane.

For the KCL data, all stacks (between 6 and 9 available) were manually reviewed for quality. These with sufficient quality were automatically masked and then reconstructed to 1.2 mm isotropic resolution 3D volumes using SVRTK [25] and aligned to standard reference space.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Previous work on inter/intra-rater variability of biometric measurements has demonstrated good reliability on SR reconstructions [3]. Inter-observer comparison was performed for each biometric measurement using the paired Wilcoxon s rank sum test and showed no significant differences ($p > 0.05$) except for TCD_cor (p-value=0.036). After correction for multiple comparisons, none of them remained significant. Analysis was performed with 2 observers that performed biometric measurements on a cohort of 26 normal fetal brains.
Furthermore, intra-rater analysis on SR reconstruction demonstrated excellent reliability with an overall intra-class correlation (ICC) reaching 0.97. Analysis was performed with 1 observer that performed the biometric measurements 3 times on a sub cohort of 5 normal fetal brains.

b) In an analogous manner, describe and quantify other relevant sources of error.

None.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

The ranking will be based on the measurement error in percentage (ME) which is the difference between estimated measurement and the actual measurement in comparison to the actual measurement.
In addition, we will compute the R^2 coefficient between the predictions and the ground truth for each region to assess the spread of the prediction, but we will not use it in the ranking.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Task 2 targets the fetal biometry accuracy of the estimated segments (in mm). It is thus a regression task and classical mean absolute/squared error metrics will be used. Also the quality of the fit between GT and predicted values will be evaluated through the R2 (but not used for the ranking).

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

This ranking system was developed in order to take two different metric types equally into account. We also wanted to determine not just average ranking, but see if algorithms performed better in when the image was high quality vs low quality, as well as how well they perform on the pathological vs the neurotypical brains.

b) Describe the method(s) used to manage submissions with missing results on test cases.

The only possibility for missing data would be if an algorithm does not find any of one label in the final label map, or if the entire label map is empty, or if biometry segments are missing or equal 0.

If there are missing results, the worst possible value will be used. For example, if a label does not exist in the label map, it will receive a DSC and VS of 0, and the HD95 will be double the max value of the other algorithms submitted (to ensure it is ranked last for that sub-ranking).

c) Justify why the described ranking scheme(s) was/were used.

The two tasks will have independent rankings. For biometry, a final score considering all cases using measurement error in percentage as a metric will be obtained for each team. For each case, each team will be ranked from low to high, where the lowest ME score receives the highest scoring rank (best). Hence, a ranking based on the sum of the ranking score across all cases can be calculated.

The results of the challenge will be run through the ChallengeR toolkit, specifically designed to calculate and display imaging challenge results [36].

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The description of missing data handling can be found above.
The mean and standard deviation of each method will be calculated using both an average of all labels as well as individually, using the ranking method as described above.

b) Justify why the described statistical method(s) was/were used.

This ranking system will take one metric into account. We also wanted to determine not just average ranking, but see if algorithms performed better in the R2, and when the image was high quality vs low quality, as well as how well they perform on the pathological vs the neurotypical brains.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Refer to Task 1.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

[1] Child mortality and causes of death. Accessed: Jan. 10, 2024. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/
[2] V. Kyriakopoulou et al., Normative biometry of the fetal brain using magnetic resonance imaging, Brain Struct. Funct., vol. 222, no. 5, 2017, doi: 10.1007/s00429-016-1342-6.
[3] M. Khawam et al., Fetal Brain Biometric Measurements on 3D Super-Resolution Reconstructed T2-Weighted MRI: An Intra- and Inter-observer Agreement Study, Front. Pediatr., vol. 9, p. 639746, Aug. 2021, doi: 10.3389/fped.2021.639746.
[4] G. Egaña-Ugrinovic, M. Sanz-Cortes, F. Figueras, N. Bargalló, and E. Gratacós, Differences in cortical development assessed by fetal MRI in late-onset intrauterine growth restriction, Am. J. Obstet. Gynecol., vol. 209, no. 2, Aug. 2013, doi: 10.1016/j.ajog.2013.04.008.
[5] A. Zugazaga Cortazar, C. Martín Martinez, C. Duran Feliubadalo, M. R. Bella Cueto, and L. Serra, Magnetic resonance imaging in the prenatal diagnosis of neural tube defects, Insights Imaging, vol. 4, no. 2, Apr. 2013, doi: 10.1007/s13244-013-0223-2.
[6] C. Clouchoux et al., Delayed cortical development in fetuses with complex congenital heart disease, Cereb. Cortex, vol. 23, no. 12, Art. no. 12, 2013.
[7] C. K. Rollins et al., Regional Brain Growth Trajectories in Fetuses with Congenital Heart Disease, Ann. Neurol., vol. 89, no. 1, Jan. 2021, doi: 10.1002/ana.25940.
[8] D. Prayer et al., MRI of normal fetal brain development, Eur. J. Radiol., vol. 57, no. 2, Feb. 2006, doi: 10.1016/j.ejrad.2005.11.020.
[9] D. A. Jarvis, C. R. Finney, and P. D. Griffiths, Normative volume measurements of the fetal intra-cranial compartments using 3D volume in utero MR imaging, Eur. Radiol., vol. 29, no. 7, Jul. 2019, doi: 10.1007/s00330-018-5938-5.
[10] T. Sanchez, O. Esteban, Y. Gomez, E. Eixarch, and M. B. Cuadra, FetMRQC: Automated Quality Control for Fetal Brain MRI, in Perinatal, Preterm and Paediatric Image Analysis, vol. 14246, D. Link-Sourani, E. Abaci Turk, C. Macgowan, J. Hutter, A. Melbourne, and R. Licandro, Eds., in Lecture Notes in Computer Science, vol. 14246. , Cham: Springer Nature Switzerland, 2023, doi: 10.1007/978-3-031-45544-5_1.
[11] T. Sanchez et al., FetMRQC: an open-source machine learning framework for multi-centric fetal brain MRI quality control. arXiv, Nov. 08, 2023. Accessed: Jan. 10, 2024. [Online]. Available: http://arxiv.org/abs/2311.04780
[12] K. Payette et al., An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal Tissue Annotation Dataset, Sci. Data, vol. 8, no. 1, Art. no. 1, Jul. 2021, doi: 10.1038/s41597-021-00946-3.

[13] K. Payette et al., Fetal brain tissue annotation and segmentation challenge results, Med. Image Anal., vol. 88, p. 102833, Aug. 2023, doi: 10.1016/j.media.2023.102833.
[14] J. Aviles Verdera et al., Reliability and Feasibility of Low-Field-Strength Fetal MRI at 0.55 T during Pregnancy, Radiology, vol. 309, no. 1, p. e223050, Oct. 2023, doi: 10.1148/radiol.223050.
[15] V. M. Campello et al., Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms; Challenge, IEEE Trans. Med. Imaging, vol. 40, no. 12, Dec. 2021, doi: 10.1109/TMI.2021.3090082.
[16] B. Glocker, R. Robinson, D. C. Castro, Q. Dou, and E. Konukoglu, Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects. arXiv, Oct. 10, 2019. Accessed: Jan. 10, 2024. [Online]. Available: http://arxiv.org/abs/1910.04597
[17] H. Lajous et al., A fetal brain magnetic resonance acquisition numerical phantom (FaBiAN), Sci. Rep., vol. 12, no. 1, 2022.
[18] L. Guibaud and A. Lacalm, Diagnostic imaging tools to elucidate decreased cephalic biometry and fetal microcephaly: a systematic analysis of the central nervous system: Editorial, Ultrasound Obstet. Gynecol., vol. 48, no. 1, Jul. 2016, doi: 10.1002/uog.15926.
[19] B. Tilea et al., Cerebral biometry in fetal magnetic resonance imaging: new reference data, Ultrasound Obstet. Gynecol., vol. 33, no. 2, Feb. 2009, doi: 10.1002/uog.6276.
[20] N. Khalili et al., Automatic brain tissue segmentation in fetal MRI using convolutional neural networks, Magn. Reson. Imaging, vol. 64, Dec. 2019, doi: 10.1016/j.mri.2019.05.020.
[21] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, Auto-Context Convolutional Neural Network (Auto-Net) for Brain Extraction in Magnetic Resonance Imaging, IEEE Trans. Med. Imaging, vol. 36, no. 11, Nov. 2017, doi: 10.1109/TMI.2017.2721362.
[22] D. Prayer et al., ISUOG Practice Guidelines: performance of fetal magnetic resonance imaging, Ultrasound Obstet. Gynecol., vol. 49, no. 5, Art. no. 5, May 2017, doi: 10.1002/uog.17412.
[23] S. Tourbier, X. Bresson, P. Hagmann, J.-P. Thiran, R. Meuli, and M. B. Cuadra, An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization, NeuroImage, vol. 118, 2015.
[24] P. Deman, S. Tourbier, R. Meuli, and M. B. Cuadra, meribach/mevislabFetalMRI: MEVISLAB MIAL Super-Resolution Reconstruction of Fetal Brain MRI v1.0. Zenodo, Jun. 05, 2020. doi: 10.5281/ZENODO.3878563.
[25] S. Tourbier, sebastientourbier/mialsuperresolutiontoolkit. Mar. 31, 2020. Accessed: Apr. 15, 2020. [Online]. Available: https://github.com/sebastientourbier/mialsuperresolutiontoolkit

[26] M. Ebner et al., An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI, NeuroImage, vol. 206, p. 116324, Feb. 2020, doi: 10.1016/j.neuroimage.2019.116324.
[27] M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, and J. A. Schnabel, Reconstruction of fetal brain MRI with intensity matching and complete outlier removal, Med. Image Anal., vol. 16, no. 8, 2012.
[28] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support, in Intraoperative Imaging and Image-Guided Therapy, F. A. Jolesz, Ed., New York, NY: Springer New York, 2014, doi: 10.1007/978-1-4614-7657-3_19.
[29] E. Schwartz et al., The Prenatal Morphomechanic Impact of Agenesis of the Corpus Callosum on Human Brain Structure and Asymmetry, Cereb. Cortex, p. bhab066, Apr. 2021, doi: 10.1093/cercor/bhab066.
[30] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images, IEEE Trans. Med. Imaging, vol. 27, no. 4, Apr. 2008, doi: 10.1109/TMI.2007.906087.
[31] C. Dong, C. C. Loy, and X. Tang, Accelerating the Super-Resolution Convolutional Neural Network, in

Computer Vision ECCV 2016, vol. 9906, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9906. , Cham: Springer International Publishing, 2016, doi: 10.1007/978-3-319-46475-6_25.
[32] A. Gholipour et al., A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth, Sci. Rep., vol. 7, no. 476, 2017, doi: 10.1038/s41598-017-00525-w.
[33] A. A. Taha and A. Hanbury, Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool, BMC Med. Imaging, vol. 15, no. 1, p. 29, Dec. 2015, doi: 10.1186/s12880-015-0068-x.
[34] L. Maier Hein et al., Metrics reloaded: Recommendations for image analysis validation. arXiv, Sep. 22, 2023. doi: 10.48550/arXiv.2206.01653.
[35] P. de Dumast, H. Kebiri, C. Atat, V. Dunet, M. Koob, and M. B. Cuadra, Segmentation of the Cortical Plate in Fetal Brain MRI with a Topological Loss, in Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis, C. H. Sudre, R. Licandro, C. Baumgartner, A. Melbourne, A. Dalca, J. Hutter, R. Tanno, E. Abaci Turk, K. Van Leemput, J. Torrents Barrena, W. M. Wells, and C. Macgowan, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, doi: 10.1007/978-3-030-87735-4_19.
[36] P. de Dumast, H. Kebiri, C. Atat, V. Dunet, M. Koob, and M. B. Cuadra, Segmentation of the cortical plate in fetal brain MRI with a topological loss, ArXiv Prepr. ArXiv201012391, 2020.
[37] V. Raina et al., Tackling Bias in the Dice Similarity Coefficient: Introducing NDSC for White Matter Lesion Segmentation, in 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia: IEEE, Apr. 2023. doi: 10.1109/ISBI53787.2023.10230755.
[38] M. Wiesenfarth et al., Methods and open-source toolkit for analyzing and visualizing challenge results, Sci. Rep., vol. 11, no. 1, p. 2369, Jan. 2021, doi: 10.1038/s41598-021-82017-6.

Further comments

Further comments from the organizers.

Note: Abbreviations of different sites are used throughout this application. Below is the legend:
Kispi: University Childrens Hospital Zurich
CHUV: Lausanne University Hospital
UCSF: University of California, San Francisco
Vienna: Medical University of Vienna
KCL: Kings College London

The paper related to FeTA challenge 2022 is currently submitted at IEEE TMI, and can be found now here https://arxiv.org/abs/2402.09463

Overview of the changes between FeTA 2022 and FeTA 2024:
-larger test dataset (in 2022, 160 testing and in 2024 180 testing)
-The new version of FeTA includes data from a fifth site KCL that uses low-field 0.55T scanner, it expands to incorporate on domain generalization towards low-field, such that the focus of the challenge is on developing segmentation methods that will work on data from different magnetic field
- includes topology errors as a metric to be used to determine the final ranking in the segmentation task
- addition of a clinically relevant task of biometry to support the analysis of MDs
- added additional evaluation metric normalize Dice to assess age/size bias