MICCAI Grand Challenge on Multi-domain Cross-time-point Infant Cerebellum MRI Segmentation 2022 : Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

MICCAI Grand Challenge on Multi-domain Cross-time-point Infant Cerebellum MRI Segmentation 2022

Challenge acronym

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

iSeg-2022

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.

Cerebellum is a rapidly developing and critical brain structure during the early postnatal stages. Cerebellar involvement has been implicated in the parthenogenesis of many neurodevelopmental disorders, e.g., autism, attention-deficit/hyperactivity disorder, and schizophrenia. Therefore, accurate segmentation of the cerebellum into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) is essential to better understand cerebellar structure and function and assist in the diagnosis and treatment of neurodevelopmental disorders. Compared with adult cerebellum, there are very few works proposed for infant cerebellum segmentation. Infant cerebellum MRIs exhibit extremely low tissue contrast and severe partial volume effects in magnetic resonance imaging (MRI), posing a huge challenge for manual and automated segmentation of the cerebellum. First, due to the low tissue contrast, the manual annotation is extremely challenging, especially for younger infant subjects (e.g., 6 months). Second, the collaborative use of multi-domain infant images (acquired from different imaging sites) makes the segmentation task more difficult. Third, there are often anatomical errors in the segmentation results. Therefore, by taking advantage of accurate manual labels from 24-month-old subjects, the aim of this challenge is to promote accurate segmentation algorithms on infant cerebellum MRIs from multiple domains.

Challenge keywords

List the primary keywords that characterize the challenge.

Infant cerebellum segmentation, multi-domain data, cross-time-point data

Year

The challenge will take place in ...

2022

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

None

Duration

How long does the challenge take?

Full 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 30+ international teams.

Publication and future plans

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

A review article will be prepared.

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.

No on-site testing.
Each team, as well as the organizers, will use their own laptops. A meeting room for 80 participants will be sufficient. A screen and a projector are needed for the presentations. A poster area is required for the poster session.

TASK: Multi-domain Cross-time-point Infant Cerebellum MRI 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.

Cerebellum is a rapidly developing and critical brain structure during the early postnatal stages. Cerebellar involvement has been implicated in the parthenogenesis of many neurodevelopmental disorders, e.g., autism, attention-deficit/hyperactivity disorder, and schizophrenia. Therefore, accurate segmentation of the cerebellum into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) is essential to better understand cerebellar structure and function and assist in the diagnosis and treatment of neurodevelopmental disorders. Compared with adult cerebellum, there are very few works proposed for infant cerebellum segmentation. Infant cerebellum MRIs exhibit extremely low tissue contrast and severe partial volume effects in magnetic resonance imaging (MRI), posing a huge challenge for manual and automated segmentation of the cerebellum. First, due to the low tissue contrast, the manual annotation is extremely challenging, especially for younger infant subjects (e.g., 6 months). Second, the collaborative use of multi-domain infant images (acquired from different imaging sites) makes the segmentation task more difficult. Third, there are often anatomical errors in the segmentation results. Therefore, by taking advantage of accurate manual labels from 24-month-old subjects, the aim of this challenge is to promote accurate segmentation algorithms on infant cerebellum MRIs from multiple domains.

Keywords

List the primary keywords that characterize the task.

Infant cerebellum segmentation, multi-domain data, cross-time-point data

ORGANIZATION

Organizers

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

Mrs. Yue Sun, (yuesun@med.unc.edu), Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA
Dr. Li Wang, (li_wang@med.unc.edu), Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA
Dr. Valerie Jewells, (valerie_jewells@med.unc.edu), Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA
Dr. Kathryn Leigh Humphreys, (k.humphreys@vanderbilt.edu), Department of Psychology and Human Development, Vanderbilt University, USA
Dr. Weili Lin, (weili_lin@med.unc.edu), Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA

b) Provide information on the primary contact person.

Mrs. Yue Sun, (yuesun@med.unc.edu)

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

Open call challenge.

Challenge venue and platform

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

MICCAI.

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

None at this moment.

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

None at this moment.

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.

Basically, no additional data and pre-trained net are allowed. The participating team can use pre-trained nets by notifying us firstly and providing the source.

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.

May participate but not eligible for awards and not listed in leaderboard.

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

We will rank the performance of team's methods in terms of each metric. We will not provide an overall rank, since the weights for different metrics are difficult to define.

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.

All performance results will be made public.

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).

We will cooperate with all participants to prepare a review paper, as we did in iSeg-2017 and iSeg-2019 Challenges.
The co-author list in the review paper is to be discussed in terms of the number of participating teams.
Participating teams can publish their own results independently without an embargo time.

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 upload the segmented images to the organizers through a link on the website.

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.

Participating teams are allowed for maximal three submissions for the testing phase, and the best result is officially counted as the challenge result.

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

Registration: April 1, 2022;
Training data release: April 1, 2022;
Testing data release: July 1, 2022;
Submission: July 1, 2022 - July 31, 2022;
Evaluation result release: August 1, 2022.

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).

The ethics approval is not necessary for the data. The data will be shared with the condition to be used only for the participation in the challenge. Their use for other scientific research will be provided by specific request. No commercial use will be granted.

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)

CC BY NC ND.

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.

Code to produce rankings will be available to get access at the challenge website.

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

Participating teams are encouraged to share their code together with the corresponding method description.

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.

There are no conflicts of interest.
Only the organizers have access to the test case 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.

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.

Multi-domain subjects at around 0 month of age.

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

Multi-domain subjects at around 6 months of age.

Imaging modality(ies)

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

T1- and T2-weighted magnetic resonance imaging (MRI) data.

Context information

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

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

The testing subjects at 6 months of age were acquired from infants at age of 150~230 days. These subjects are excluded from this challenge if they 1) have a first degree relative with autism, intellectual disability, schizophrenia, or bipolar disorder, 2) have any significant medical and/or genetic conditions affecting growth, development, or cognition, or 3) have any contraindication to MRI.

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

N/A.

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.

Cerebellum shown in MRI data.

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.

Label each brain voxel as white matter, gray matter or cerebrospinal fluid for cerebellum in multi-domain MRIs.

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.

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).

3T Siemens Prisma scanner, 3T GE scanner, and 3T Philips scanner

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

Scanner: Modality, TR/TE (ms), Head Coil, Resolution (mm3)
1. 3T Siemens: T1w, 2400/2.2, 32-channel, 0.8×0.8×0.8;
T2w, 3200/564, 32-channel, 0.8×0.8×0.8.
2. 3T GE: T1w, 7.6/2.9, 32-channel, 0.9×0.9×0.8;
T2w, 25.2/91.4, 32-channel, 1.0×1.0×1.0.
3. 3T Philips: T1w, 10/4.6, N/A, 1.0×1.0×1.0;
T2w, 2500/310, N/A, 0.8×0.8×0.8.

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.

1. Subjects scanned with Siemens scanner are acquired from the site of Biomedical Research Imaging Center (BRIC) at the University of North Carolina at Chapel Hill and the Center for Magnetic Resonance Research (CMRR) at the University of Minnesota (UNC/UMN (BCP)).
2. Subjects scanned with a GT scanner are acquired from Stanford University.
3. Subjects scanned with a Philips scanner are acquired from Vanderbilt University.

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).

Subjects are acquired with a team of radiographers, research nurses and clinical fellows as appropriate.

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.

Training and test cases both represent T1- and T2-weighted MRIs of cerebellum regions with full annotations (voxel level). The training and validation cases are acquired from infants at 24 months of age from UNC/UMN (BCP), while the test cases are acquired from infants at 6 months of age from UNC/UMN (BCP), Stanford University and Vanderbilt University. Note that all subjects are cross-sectional.

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

13 training subjects randomly selected from UNC/UMN (BCP) (24 months);
5 testing subjects randomly selected from UNC/UMN (BCP) (24 months);
15 testing subjects randomly selected from UNC/UMN (BCP), Stanford University and Vanderbilt University (6 months).

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

The number of training subjects from the UNC/UMN (BCP) is 18 (24 months), and the number of testing subjects from UNC/UMN (BCP), Stanford University, and Vanderbilt University is 10 (5 at 24 months, 5 at 6 months), 5 (6 months), and 5 (6 months), respectively. The reason for the different numbers from different sites is that more 24-month-old subjects have been collected in the UNC/UMN (BCP) while fewer subjects are available from the testing sites. On the other hand, comparing with 24-month-old subjects, the 6-month-old subject shows lower tissue contrast and extremely folded tiny structures, making annotation editing becomes increasingly difficult and time-consuming. Therefore, only a limited number of 6-month-old subjects are manually annotated as testing data for quantitative comparisons.

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.

The gender distribution of subjects is similar to the real-world distribution: 50% male and 50% female.

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.

We used an anatomy-guided densely-connected U-Net [1] to generate an initial segmentation. The initial segmentations were later followed by manual correction under the guidance of an experienced neuroradiologist (Dr. Valerie Jewells, UNC-Chapel Hill).

1. L. Wang, G. Li, F. Shi, X. Cao, C. Lian, D. Nie, M. Liu, H. Zhang, Z. Wu, W. Lin, and D. Shen, "Volume-based analysis of 6-month- old infant brain MRI for autism biomarker identification and early diagnosis," in MICCAI, vol. 11072, 2018, pp. 411–419.

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.

N/A.

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.

The manual correction was performed under the guidance of an experienced neuroradiologist (Dr. Valerie Jewells, UNC-Chapel Hill).

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.

Only one annotation for each subject.

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.

The resolution of all images was resampled into 0.8×0.8×0.8 mm3, and T2w images were linearly aligned with their corresponding T1w images. We performed skull stripping and extraction of the cerebellum by leveraging an infant-dedicated pipeline (i.e., iBEAT V2.0 Cloud http://www.ibeat.cloud). In this challenge, we choose imaging data without motion artifacts as training and testing set, and the data is not 3D reconstructed.

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.

Anatomical errors due to extremely low tissue contrast and severe partial volume effect.

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

N/A.

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)

Dice Similarity Coefficient (DSC);
95% Hausdorff distance (HD95).

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

Dice Similarity Coefficient (DSC) are commonly used in segmentation assessment [1].
95% Hausdorff distance as opposed to standard HD: Try to avoid that outliers have too much weight.

1. http://neobrains12.isi.uu.nl/evaluation.php

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.

Similar with iSeg-2017 and iSeg-2019, we will only provide a rank for each metric. We will not provide an overall rank, since the weights for different metrics are difficult to define.

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

Missing subjects are not allowed, and any missing tissue structure in the testing results would result in no ranking of the team.

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

In terms of volume calculation, Dice Similarity Coefficient (DSC) is more important than 95% Hausdorff distance (HD95). However, in terms of cortical thickness estimation, 95% Hausdorff distance (HD95) is more important. Another reason is that we want to attract more teams to participate and we think each team who participates is a winner.

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.

Wilcoxon signed-rank test used for paired samples;
Wilcoxon rank-sum test used for unpaired samples.

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

N/A.

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.

N/A.

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. L. Wang, D. Nie, and G. Li et al., "Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iSeg-2017 Challenge," IEEE Transactions on Medical Imaging, vol. 38, no. 9, pp. 2219-2230, Feb 27 2019.
2. Y. Sun, K. Gao, and Z. Wu et al., “Multi-site infant brain segmentation algorithms: The iSeg-2019 Challenge,” IEEE Transactions on Medical Imaging, vol. 40, no. 5, pp. 1363–1376, 2021.
3. B. Howell, M. Styner, and W. Gao et al., “The UNC/UMN baby connectome project (BCP): An overview of the study design and protocol development,” NeuroImage, vol. 185, pp. 891–905, Jan. 2019.

Further comments

Further comments from the organizers.

N/A.