Intrapartum Ultrasound Grand Challenge 2024: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Intrapartum Ultrasound Grand Challenge 2024

Challenge acronym

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

IUGC2024

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.

Over the years, global caesarian section(CS) rates have significantly increased from around 7% in 1990 to 21% today, surpassing the ideal acceptable CS rate which is around 10% to 15%, according to the World Health Organization (WHO). These trends are projected to continue increasing over the current decade, where both unmet needs and overuse are expected to coexist with the projected global rate of 29% by 2030. In light of this and concerns regarding adverse health and economic consequences following operative birth, there is increasing recognition that prevention of avoidable cesarean births is important, provided it does not increase rates of adverse neonatal or maternal outcomes.

Active management of labour has been proposed as a means of reducing unnecessary CSs. The WHO also issued guidelines on intrapartum care, which include a strong recommendation in favor of using a modified partograph and a recommendation of digital vaginal examination (VE) of the fetal head (FH) station every four hours during the first stage of labor. FH station is the level of the FH in the birth canal in relation to the imaginary line between the maternal ischial spines. The FH is considered to be engaged at station 0 when the leading part of the skull reaches this imaginary line. VE of the FH station has been demonstrated to be subjective, with a 30% to 34% error in classifying the FH station as highpelvis, midpelvis, lowpelvis, or outlet on a birth simulator with sensors by clinicians with varying experience. The majority of errors were caused by misdiagnosis of a midpelvic station for a true high-pelvic station, which, in real life, could mislead clinicians to perform a midcavity instrumental vaginal delivery when cesarean delivery might have been a better choice.

Accurate assessment of the FH station is crucial to ensure that childbirth proceeds normally, allowing for the early detection of any deviations from the anticipated course, enabling timely intervention to mitigate potential maternal or fetal complications. Compared with VE, an ultrasound examination, which allows visualization of the fetal structures and their relationship with maternal structures, is quite straightforward, and provides more objective and accurate results in assessing the FH station. It is neither time consuming nor causes discomfort of

patient. The International Society of Ultrasound in Obstetrics and Gynecology has issued practical guidelines on intrapartum ultrasound in 2018 and recommended that an ultrasound assessment should be conducted when there is suspected delay or arrest of the first or second stage of labor or before considering assisted vaginal delivery [1]. FH station is assessed by transperineal ultrasound using the maternal pubic symphysis (PS) or the perineum as landmark for reference of measurement. Of the various ultrasound parameters, angle of progression (AOP) and head symphysis distance (HSD) have been suggested as the reliable sonographic parameters to predict the outcome of the instrumental vaginal delivery. AOP is the angle between the long axis of the pubic bone and a line from the lowest edge of the pubis drawn tangentially to the deepest bony part of the fetal skull, whereas HSD is the distance between the lowermost edge of the pubic symphysis and the nearest point of the fetal head along a line perpendicular to the long axis of the pubic symphysis.

Despite its clear benefits, the implementation of intrapartum ultrasound in labor and delivery presents technical challenges. It is particularly worth mentioning that, different from antepartum ultrasound for professional sonographer in most obstetric examinations, intrapartum ultrasound is a relatively new technology for non ultrasound trained professionals to provide continuous 24h, on site, on demand labor and delivery services at the bedside. Moreover, labor is a dynamic process and multiple ultrasound examinations are done for assessing changes in FH station. Diagnostic accuracy is severely affected by subjective factors such as physician experience and fatigue. Undoubtedly, the availability of immediate and convenient diagnostic and intervention guided ultrasound support is required in the demanding labor and delivery environment.

The biomedical impact of this challenge is profound. Accurate and timely assessments can significantly reduce unnecessary CS rates, leading to better health outcomes for mothers and babies. Technically, this challenge calls for the development of automatic, user friendly systems for fetal biometrics, aiming to minimize intra and inter observer variability and enhance the reliability of measurements [2-7]. Such advancements could revolutionize labor management, blending the precision of technology with the nuances of human care.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Intrapartum Ultrasound, Artificial Intelligence, Fetal Station, Fetal Biometry, Angle of Progression, Head Symphysis Distance

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.

none

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.

Our PSFHS challenge (MICCAI 2023) attracted 187 participants, so for IUGC 2024 we hope to have more than 50 teams participating.

Publication and future plans

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

We aim to summarize the design, proposed methods and results of the challenge in a manuscript to be submitted to a peer-reviewed scientific journal in medical image analysis. To this end, we invite the participants to contribute by describing their methods and experiences and submitting their manuscripts according to the guidelines (https://conferences.miccai.org/2024/en/PAPER-SUBMISSION-AND-REBUTTAL-GUIDELINES.html). Furthermore, we aim to make the code of the best-performing methods publicly available for the purpose of reproducing results and further research.

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.

For algorithm implementation and training, the participants will use their own resources. For testing as part of the challenge, we would use the platform CodaLab.

TASK 1: 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.

Over the years, global caesarian section(CS) rates have significantly increased from around 7% in 1990 to 21% today surpassing the ideal acceptable CS rate which is around 10% to 15% according to the World Health Organization (WHO). These trends are projected to continue increasing over the current decade where both unmet needs and overuse are expected to coexist with the projected global rate of 29% by 2030. In light of this and concerns regarding adverse health and economic consequences following operative birth, there is increasing recognition that prevention of avoidable cesarean births is important, provided it does not increase rates of adverse neonatal or maternal outcome.

Active management of labour has been proposed as a means of reducing unnecessary CSs. The WHO also issued guidelines on intrapartum care, which include a strong recommendation in favor of using a modified partograph and a recommendation of digital vaginal examination (VE) of the fetal head (FH) station every four hours during the first stage of labor. FH station is the level of the FH in the birth canal in relation to the imaginary line between the maternal ischial spines. The FH is considered to be engaged at station 0 when the leading part of the skull reaches this imaginary line. VE of the FH station has been demonstrated to be subjective, with a 30% to 34% error in classifying the FH station as highpelvis, midpelvis, lowpelvis, or outlet on a birth simulator with sensors by clinicians with varying experience. The majority of errors were caused by misdiagnosis of a midpelvic station for a true high-pelvic station, which, in real life, could mislead clinicians to perform a midcavity instrumental vaginal delivery when cesarean delivery might have been a better choice.

Accurate assessment of the FH station is crucial to ensure that childbirth proceeds normally, allowing for the early detection of any deviations from the anticipated course, enabling timely intervention to mitigate potential maternal or fetal complications. Compared with VE, an ultrasound examination, which allows visualization of the fetal structures and their relationship with maternal structures, is quite straightforward, and provides more objective and accurate results in assessing the FH station. It is neither time consuming nor causes discomfort of patient. The International Society of Ultrasound in Obstetrics and Gynecology has issued practical guidelines on intrapartum ultrasound in 2018 and recommended that an ultrasound assessment should be conducted when there is suspected delay or arrest of the first or second stage of labor or before considering assisted vaginal delivery [1]. FH station is assessed by transperineal ultrasound using the maternal pubic symphysis (PS) or the perineum as landmark for reference of measurement. Of the various ultrasound parameters, angle of progression (AOP) and head symphysis distance (HSD) have been suggested as the reliable sonographic parameters to predict the outcome of the instrumental vaginal delivery. AOP is the angle between the long axis of the pubic bone and a line from the lowest edge of the pubis drawn tangentially to the deepest bony part of the fetal skull, whereas HSD is the distance between the lowermost edge of the pubic symphysis and the nearest point of the fetal head along a line perpendicular to the long axis of the pubic symphysis.

Despite its clear benefits, the implementation of intrapartum ultrasound in labor and delivery presents technical challenges. It is particularly worth mentioning that, different from antepartum ultrasound for professional sonographer in most obstetric examinations, intrapartum ultrasound is a relatively new technology for non ultrasound trained professionals to provide continuous 24h, on site, on demand labor and delivery services at the bedside. Moreover, labor is a dynamic process and multiple ultrasound examinations are done for assessing changes in FH station. Diagnostic accuracy is severely affected by subjective factors such as physician experience and fatigue. Undoubtedly, the availability of immediate and convenient diagnostic and intervention guided ultrasound support is required in the demanding labor and delivery environment.

The biomedical impact of this challenge is profound. Accurate and timely assessments can significantly reduce unnecessary CS rates, leading to better health outcomes for mothers and babies. Technically, this challenge calls for the development of automatic, user friendly systems for fetal biometrics, aiming to minimize intra and inter observer variability and enhance the reliability of measurements [2-7]. Such advancements could revolutionize labor management, blending the precision of technology with the nuances of human care.

Keywords

List the primary keywords that characterize the task.

Intrapartum Ultrasound, Artificial Intelligence, Fetal Station, Fetal Biometry, Angle of Progression, Head Symphysis Distance

ORGANIZATION

Organizers

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

Jieyun Bai, Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Information Technology, Jinan University, CN
Karim Lekadir, Artificial Intelligence in Medicine Lab (BCN-AIM), Barcelona, Spain
Dong Ni, Shenzhen University, CN
Saad Slimani, Ibn Rochd University Hospital, Hassan II University, Casablanca, Morocco
Campello Roman Victor Manuel, Artificial Intelligence in Medicine Lab (BCN-AIM), Barcelona, Spain
Benard Ohene-Botwe, Department of Radiography, School of Biomedical and Allied Health Sciences, College of Health Sciences, University of Ghana, Accra, Ghana.
Yaosheng Lu, Jinan University, Guangzhou, CN
Gaowen Chen, Zhujiang Hospital of Southern Medical University, Guangzhou, CN
Hongying Hou, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, CN
Di Qiu, the First Affiliated Hospital of Jinan University, Guangzhou, CN
Zihao Zhou, Jinan University, Guangzhou, CN

b) Provide information on the primary contact person.

Jieyun Bai (jbai996@aucklanduni.ac.nz)

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

Repeated event with annual fixed submission deadline

Challenge venue and platform

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

MICCAI (a 2-hour session)

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

https://competitions.codalab.org/ (will setup once the proposal is accepted)

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.

Be restricted to the data provided by the challenge.

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.

Everyone is encouraged to enter the competition but the conflict of interests should be stated clearly by participants for paper reviewing.

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

Certificates will be provided for the top 10 performing teams. We are actively seeking sponsorship and we anticipate being able to provide cash prizes. Our first MICCAI challenge (PSFHS2023) was sponsored by Guangzhou Lian-Med Technologies Co., Ltd., who provided the prizes for the winning team, so we do not anticipate any issues sourcing prizes for 2024. Note: A prerequisite for receiving cash prizes and award certificates is that participants need to provide the source code, give a presentation about their methods, as well as submit conference papers describing methods and results in accordance with the guidelines (https://conferences.miccai.org/2024/en/PAPER-SUBMISSION-AND-REBUTTAL-GUIDELINES.html).

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, all participants are invited to submit conference papers, and the top 10 that meet the winning conditions will be invited to participate in a challenge event to showcase their work.

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).

All the participating teams can submit their workshop papers. These submissions will be reviewed by 3 reviewers. In addition, participating teams can also anyway publish their own results independently after an embargo time of 6 months. We also intend to coordinate with the top 10 teams to submit a journal article summarizing the main results and conclusions drawn from the challenge.

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.

For the purpose of result verification and to encourage reproducibility and transparency, all entries must submit the following:
-Mask images indicating image id, pixel-wise label for background (0), fetal head region (1) and pubic symphysis (2), AOP and HSD. This is to ensure that all submissions are fairly and correctly evaluated for comparisons.
-A paper highlighting the contribution of the submission, but not limited to, the method, experimental results and analysis, prepared according to the format stipulated by MICCAI 2024. All challenge entries should be accompanied by a description of the method.
-GitHub repository URL containing all source codes for their implemented method and all other relevant files such as feature/parameter data. To help publicize our workshop and domain area, please mention (or add relevant links to) IUGC 2024 and MICCAI 2024. The participants may provide this URL in a simple text file while submitting.
-For all files, participants should submit a single zip file and upload it to the submission system as supplementary material.
The submission link will be made available starting 01/08/2024.

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.

The participating teams will be able to validate their results based on the validation set provided by the organizers. Submissions to IUGC 2024 are issued a validation score. This is to provide a sanity check of the submission (ensure the submission is in the correct format) and is not intended to be used for algorithm ranking or evaluation.

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: 15/03/2024
Registration Starting: 01/04/2024
Submission deadline: 01/08/2024
Winner and invitation speakers: 01/09/2024
Announcement of results at MICCAI 2024: 10/10/2024 (subject to change depending on the MICCAI 2024 deadlines)

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).

All data is anonymized. Ethics approval by the Medical Ethics Committee of the First Affiliated Hospital of Jinan University (No. JNUKY-2022-019), Zhujiang Hospital of Southern Medical University (No. 2023-SYJS-023) and the Third Affiliated Hospital of Sun Yat-sen University (No. [2021]20-367).

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 (Attribution-NonCommercial-NoDerivatives). To download the dataset, please complete the agreement form (see supplementary form) and forward it to the dataset organizer (Jieyun Bai at jbai996@aucklanduni.ac.nz)

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.

We will provide an online platform (https://competitions.codalab.org/) to evaluate the results. For transparency, we will release the source code used for calculating final scores after the closing date of the challenge.

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

Algorithm code release and corresponding conference paper submission will be prerequisites for award eligibility and further use of the data. To this end, GitHub repository URL containing the source code for their implemented method and conference paper following the guidelines (https://conferences.miccai.org/2024/en/PAPER-SUBMISSION-AND-REBUTTAL-GUIDELINES.html) will need to be provided.

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. This research topic is supported by the Guangzhou Science and Technology Planning Project (2023B03J1297). Guangzhou Lian-Med Technologies Co., Ltd. is the principal sponsor of the challenge by collecting and providing clinical data.
Only the organizers and technical groups of the challenge have access to 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

Assistance, Training, Diagnosis, Research, CAD, Decision support, Screening.

Task category(ies)

State the task category(ies)

Examples:

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

Classification, Segmentation, Prediction

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 consists of pregnant women required to assess fetal station in labor.

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

Pregnant women required to assess fetal station in labor.

Imaging modality(ies)

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

2D ultrasound imaging

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 parameters (i.e., AOP and HSD) were measured with these ultrasound images with the maternal pubic symphysis (PS) and fetal head (FH). AOP is the angle between the long axis of the pubic bone and a line from the lowest edge of the pubis drawn tangentially to the deepest bony part of the fetal skull, whereas HSD is the distance between the lowermost edge of the pubic symphysis and the nearest point of the fetal head along a line perpendicular to the long axis of the pubic symphysis.

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

Pregnant women with a variety of age (from 18-year old to 46-year old).

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.

Maternal pubic symphysis (PS) and fetal head (FH) shown in ultrasound video 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.

The target structures of the to be developed algorithms are contours of maternal pubic symphysis (PS) and fetal head (FH).

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, Specificity
Additional points: Aim 1.) Accurate detection of standard intrapartum ultrasound planes within ultrasound video data. This aim assesses the efficacy of computer algorithms in classifying images, specifically identifying ultrasound images suitable for biometric parameter measurement.
Aim 2.) Accurate segmentation of maternal pubic symphysis (PS) and fetal head (FH) in ultrasound images. This aim evaluates the proficiency of computer algorithms in image segmentation, focusing on accurately delineating target contours to minimize false-positive identification of anatomical structures.
Aim 3.) Automatical and accurate measurement of biometric parameters in accordance with clinical guidelines. Specifically, this involves measuring the Angle of Progression (AOP) and Head Symphysis Distance (HSD), with AOP being the angle between the pubic bone's long axis and a line tangential to the fetal skull's deepest bony part from the lowest edge of the pubis, and HSD being the distance from the lowermost edge of the pubic symphysis

to the nearest fetal head point, along a line perpendicular to the pubic symphysis's long axis.

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).

Transperineal ultrasound examinations were performed in standard B-mode ultrasound using systems of different vendors such as Voluson P8, Esaote Mylab and Lian-med ObEye.

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

In order to obtain high-quality images, the transducer was prepped by covering it with a surgical latex glove filled with coupling gel, then the prepped transducer, after applying gel, was placed between labia below the pubic symphysis to obtain a sagittal plane, small adjustments in the form of lateral movements of the probe were made until an image obtained showed clear maternal pelvic (pubic symphysis) and fetal (fetal skull) landmarks that did not show any shadows from the pubic rami.

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.

Data were acquired from three university hospitals (the First Affiliated Hospital of Jinan University, Zhujiang Hospital of Southern Medical University, and the Third Affiliated Hospital of Sun Yat-sen 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).

Data were acquired by specialized teams consisting of sonographers, obstetricians, and technologists, both with more than seven years professional experience.
Manual segmentation and measurement were performed by three sonographers with experience in ultrasound imaging.

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.

A case (training or test case) consists of one ultrasound video, one or several ultrasound standard planes, one mask of manually segmented maternal pubic symphysis (PS) and fetal head (FH), and corresponding AOP and HSD.

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

Training cases: 700
Test cases: 300

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

The number of training cases is a trade-off between effort of manual classification, segmentation and measurement and necessity for sufficient training data. To our knowledge, this is the largest publicly available labeled intrapartum ultrasound video dataset.

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.

All training cases (n=700) are drawn from one hospital (Zhujiang Hospital of Southern Medical University, Esaote Mylab)
Test cases are split into three subgroups: 50 are drawn from the same hospital as the training cases (Zhujiang Hospital of Southern Medical University, Esaote Mylab), 200 are drawn from the First Affiliated Hospital of Jinan University (Lian-med ObEye), and 50 are drawn from the Third Affiliated Hospital of Sun Yat-sen University (Voluson P8).
The rationale behind this selection of test cases is to assess the performance of algorithms on data with slightly different characteristics and thus estimate algorithm robustness and generalizability.

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.

Manual segmentation and measurement were performed by three sonographers with experience in ultrasound imaging. To this end, maternal pubic symphysis (PS) and fetal head (FH) were manually segmented on the ultrasound image data using the software Pair.

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 following annotation protocol was defined:
Step 1: Identification of intrapartum ultrasound standard planes by selecting high-quality images containing intact targets (i.e., maternal pubic symphysis (PS) and fetal head (FH)) from an ultrasound video;
Step 2: Manual free-hand segmentation of maternal pubic symphysis (PS) and fetal head (FH);
Step 3: Manual measurement of biometric parameters (AOP and HSD) as defined by clinical guidelines: AOP is the angle between the long axis of the pubic bone and a line from the lowest edge of the pubis drawn tangentially to the deepest bony part of the fetal skull, whereas HSD is the distance between the lowermost edge of the pubic symphysis and the nearest point of the fetal head along a line perpendicular to the long axis of the pubic

symphysis.

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.

Three experts annotated training and test data:
At the First Affiliated Hospital of Jinan University, an obstetrician with seven years of experience in intrapartum ultrasound examinations;
At the Zhujiang Hospital of Southern Medical University, a sonographer with ten years of experience in ultrasound imaging and experience in machine learning research annotated all data;
At the Third Affiliated Hospital of Sun Yat-sen University, a sonographer with eight years of experience in intrapartum ultrasound examinations.

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.

The US images were anonymized by first removing any patient-related information on each image. Then all images were renamed and converted to the same format.

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.

The possible error sources related to the image annotation include intra-rater variability. For inter-rater variability, we have each of the annotators evaluate 50 videos from each of the other two raters' sets. This means each annotator will assess an additional 100 videos not originally annotated by them. In total, this adds 150 videos per annotator for assessment, sourced equally from the test datasets of Jinan University Hospital, Sun Yat-sen University Hospital, and Zhujiang Hospital.

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

The delineation of structures in ultrasound images is a challenging task, as some of the boundaries are less well-defined.

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)

1) Accuracy, F1-score, Matthew's Correlation Coefficient and Area under curve for image classification;
2) Dice similarity coefficient, Hausdorff distance, average symmetric surface distance for target segmentation;
3) The difference between predicted and manually measured ultrasound parameters.

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

Classification metrics: Area under curve represents the graphical representation of the model discriminability between fetal US image classes, computed using the estimated true positive and false positive rates. Accuracy is the most common and accessible metric to evaluate the performance of a binary classifier. What it does is tell us how many times our model has correctly classified an item in our dataset with respect to the total. The F1 score and the Matthews Correlation Coefficient (MCC) are both metrics commonly used in binary classification scenarios.
Segmentation metrics: The dice similarity coefficient is more sensitive to the inner filling of the mask, while Hausdorff distance or average symmetric surface distance is more sensitive to the segmented boundary. The combination of these three scores is sensible for our application.
The absolute value of the difference between predicted and manually measured ultrasound parameters indicates whether the prediction is consistent with the label.

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.

Step 1: Separate rankings will be computed based on each metric;
Step 2: From the nine ranking tables, the mean ranking of each participant will be computed as the numerical mean;
Step 3: In case of equal ranking, the achieved Hausdorff distance, average symmetric surface distance, AOP and HSD will be used as a tiebreak.of the single rankings;

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

Submissions with missing results on test cases will not be considered for the leaderboard.

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

The ranking scheme was selected based on the robustness analysis capabilities provided by the challengeR package. This decision was driven by the need to ensure the reliability and fairness of the competition rankings. The challengeR package utilizes Kendall's tau as its core metric for evaluating rank correlation, which effectively measures the degree of similarity between different sets of rankings. A significant aspect of employing Kendall's tau is its ability to identify consistent ranking patterns, thereby affirming the stability and consistency of the results amidst potential variations. The implementation of this measure is crucial in preserving the integrity of the competition, as it guarantees that the rankings reflect a fair and accurate assessment of participant performance, thus maintaining the trust and confidence of all involved parties.

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.

For each submission, mean, standard deviation over the test cohort and range of the defined metrics will be computed. In addition, these statistics will be calculated separately for the three parts of the test data set (data from the three Hospitals).

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

These statistical will be used in order to provide an overview of algorithm performance and reliability as well as generalizability across institutions.

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.

The challenge organizers will analyze the submitted algorithms in a journal publication. Depending upon the submitted methods, this may include comparisons of different classes of algorithms (e.g., deep learning vs others), typical failure modes, etc.

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]Ghi T, Eggebø T, Lees C, et al. ISUOG Practice Guidelines: intrapartum ultrasound[J]. Ultrasound in Obstetrics & Gynecology, 2018, 52(1): 128-139.
[2]https://ps-fh-aop-2023.grand-challenge.org/
[3]Zhou M, Yuan C, Chen Z, et al. Automatic Angle of Progress Measurement of Intrapartum Transperineal Ultrasound Image with Deep Learning[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2020: 406-414.
[4]Lu Y, Zhou M, Zhi D, et al. The JNU-IFM dataset for segmenting pubic symphysis-fetal head[J]. Data in brief, 2022, 41: 107904.
[5]Chen Z, Ou Z, Lu Y, et al. Direction-guided and multi-scale feature screening for fetal head-pubic symphysis segmentation and angle of progression calculation[J]. Expert Systems with Applications, 2023: 123096.
[6]Lu Y, Zhi D, Zhou M, et al. Multitask deep neural network for the fully automatic measurement of the angle of progression[J]. Computational and Mathematical Methods in Medicine, 2022, 2022.
[7]Bai J, Sun Z, Yu S, et al. A framework for computing angle of progression from transperineal ultrasound images for evaluating fetal head descent using a novel double branch network[J]. Frontiers in Physiology, 2022, 13: 2565.

Further comments

Further comments from the organizers.

We'd like to highlight a key distinction: Unlike antepartum ultrasound, typically performed by professional sonographers during obstetric examinations, intrapartum ultrasound represents a novel approach. This technology enables non-ultrasound-trained professionals to offer continuous, on-demand labor and delivery services bedside, around the clock. Labor being a dynamic process necessitates multiple ultrasound scans to monitor changes in fetal station. This aspect sets our current focus apart from the 2023 challenge, which was centered on ultrasound image segmentation. This year, we are delving into the clinical application, emphasizing multi-parameter automatic measurements derived from ultrasound videos. We are excited about this direction and eagerly anticipate your support.