Towards real world medical image analysis: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Towards real world medical image analysis

Challenge acronym

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

CARE (CARE - Comprehensive analysis & computing of real-world medical images)

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.

Note that:
The updates are marked by ***(Update)***. For convenience, reviewers can search "(Update)".

Many foundation models for medical image analysis, such as Segment Anything Model (SAM), have been released and proved to be useful in multiple tasks. However, their effectiveness on real world medical imaging data has not been explored. For example, specific images targeted on organs with large deformation, i.e., heart and liver, exhibit greater challenges for analysis.
[challenges]
First, the misalignment caused by the respiratory motion and cardiac pulsation increases the complexity of performing joint analysis on these data. Second, the inhomogeneity of real-world medical images poses a challenge, including the diversity of modality and the distribution shift caused by collection from various centers. Third, it could be more challenging for those foundation model to work on irregular ROIs, such as lesions or scars, whose size can be very small and shape irregular. Hence, developing effective and efficient transfer learning approaches to fully utilize those foundation models for real world medical image segmentation is of great values.

[our contribution]
In this challenge, we set up a fair and public stage for developing and validating algorithms and applications of transferring foundation models to diverse real-world medical images to address specific practical medical image analysis problems, as well as using conventional methods without pre-trained models.

[Datasets]
Four specific datasets will be released grouped by clinical requirements, targeted on organs with grand large deformation, i.e., heart and liver, and consisting of over 1250 patients from three continents. The diversity of datasets manifests in the following aspects, i.e., multi-continents: collected from over 18 centers across three

continents, multi-modality: various modalities are encompassed, misalignment: inherent misalignment exists caused by the respiratory motion and cardiac pulsation, and missing data: refers to the modality missing occurred in practice.

[Track]
Five tracks will be held in this challenge, including one comprehensive issue and four specific tracks with corresponding images and clinical problems, namely, (1) Transferring Foundation model track, (2) MyoPS ++, (3) LiFS, (4) Whole Heart Segmentation ++, (5) LAScarQS++. Specifically, the first track aims at a uniform Transferring Foundation model for the generality across the other four tracks, namely to address all or partial tasks. The uniform model will undergo comprehensive evaluation, with various metrics being integrated for ranking purposes.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Foundation Model, Generalizability, Real-World medical image analysis, Myocardium pathology segmentation, Liver fibrosis staging and segmentation, Whole heart segmentation, Left atrial and scar quantification and segmentation

Year

2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

N/A

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.

100

Publication and future plans

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

1. We may consider to submit a proceeding.
2. We may consider to submit a benchmark 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.

1 projector, 1 computer (or 1 laptop), 1 monitor, 2 loud speakers, 3 microphones.

TASK 1: Transferring Foundation Models for Multi-continent Real-World Medical Image Analysis

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.

While foundation models like the Segment Anything Model (SAM) have demonstrated efficacy in various medical image analysis tasks, their performance on real-world data remains underexplored. Specifically, these models are typically trained for normal and large targets such as the liver and lungs, yet real-world data often originates from different centers with diverse modalities. Furthermore, the application of foundation models to complicated Regions of Interest (ROIs), like lesions or scars, poses additional challenges due to their small size and irregular shapes. This task aims to address these gaps for the development of effective and efficient transfer learning approaches to maximize the utility of foundation models in the segmentation of real-world medical images.
***(Update) This challenge is focused on transferring foundation models for the segmentation of all the regions of interest (ROIs) across the following tracks. Each of the following tracks concentrates on the segmentation of specific medical targets with corresponding image data. Moreover, all the image data provided in this challenge can be utilized for this track, and competitors are advised to utilize any additional public datasets they find relevant. The transferring from all foundational models is permitted, encompassing SAM, FastSAM, MobileSAM, and MedSAM, etc.
***

Keywords

List the primary keywords that characterize the task.

Foundation Model, Real-World segmentation, Transferability, Generalizability

ORGANIZATION

Organizers

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

Fuping Wu, Department of Population Health, University of Oxford, Oxford, UK.
Shangqi Gao, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK.
Xiahai Zhuang, School of Data Science, Fudan University, Shanghai, China.

b) Provide information on the primary contact person.

Fuping Wu: fuping_wu@outlook.com, Shangqi Gao: 18110980005@fudan.edu.cn

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 opens for new submissions after conference deadline)

Challenge venue and platform

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

None.

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

https://cmt3.research.microsoft.com/CARE_2024

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

https://zmiclab.github.io/projects/CARE_2024/

Participation policies

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

only automatic methods

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.

Two settings are available.
Publicly available data is allowed or No additional data allowed.
(Update) Publicly available data is allowed or No additional data is allowed. Specifically, two separate rankings for these two groups will be established: (1) additional data used; (2) only use the training data from this challenge without any additional data.

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 not participate.

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

One best paper and two runner-ups.

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.

Participating teams can choose whether the 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).

1. The participating teams may publish their own results separately and are encouraged to cite our challenge.
2. There is 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.

Algorithm output may be evaluated online or be sent by e-mail.
Similar to LAScarQS2022, we established an online evaluation platform to support outputs online evaluation.

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.

(Update)
Maximum number of model evaluations on the validation set: 3 each day after the submission platform opens.
Maximum number of final model evaluations on the test set: 1 in total.

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: April 1st, 2024
***(Updata)Validation evaluation: May 1st, 2024
Test data evaluation: July 1st, 2024***
Submission deadline: July 15th, 2024
Notification of acceptance: July 23rd, 2024
Workshop camera ready: July 31st, 2024
Workshop date: (TBA)

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).

No ethics approval is needed.
The study had been approved by the institutional review board and the data had been anonymized.

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-SA

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.

The source code of evaluation won't be available. However, we will provide an online evaluation platform for long-term research purposes.

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

Participants are encouraged to publish their code.

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.

The sponsors are not decided yet.
Only organizers can access the test data 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

Longitudinal study, Treatment planning, Decision support, Diagnosis, CAD.

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.

(Update) The target cohorts of the task include all segmentation targets among the following four tracks, namely, segment scars and edema of Myo in Task 2, liver structure in Task 3, seven substructures of the heart in Task 4, and LA cavity and scar regions in Task 5.

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

The challenge cohort is same as the target cohort.

Imaging modality(ies)

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

It is the collection of the following Task 2-5.

Context information

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

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

No additional information.

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

None

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.

Including those regions in Tasks 2-5 as described below, namely, LV, Liver region, whole heart structure, and LA cavity and scars.

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.

(Update) The algorithm targets include all segmentation targets among the following tasks, namely, scars and edema of Myo in Task 2, liver structure in Task 3, seven substructures of the heart in Task 4, and LA cavity and scar regions in Task 5.

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, Robustness, Generalizability, Transferability.

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).

As Task1 employs all data in Task 2-5, the data acquisition devices are the same as those described below for Task 2-5.

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

Task1 employs all data in Task 2-5, and those data acquisition details can be found below for each task.

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.

(Update)
Data utilized for the challenge are collected from 10 institutes with 18 healthcare centers across three continents. The list of institutes and (if any) specific healthcare centers is as follows,
1. Inserm (France, Europe): Centre Hospitalier Régional Universitaire (CHRU) de Nancy,Centre Hospitalier Universitaire (CHU) de Bordeaux, Hôpital européen Georges-Pompidou (AP-HP HEGP) Paris, Hôpitaux Universitaires Pitié Salpêtrière (AP-HP La Pitié-Salpêtrière) Paris, Centre hospitalier universitaire (CHU) de Toulouse
2. King's College London (UK, Europe): St Thomas' Hospital London; Guy's Hospital; Imaging Sciences, Royal Brompton Hospital
3. Sheffield Teaching Hospital (UK, Europe)
4. Fudan University (China, Asia): Zhongshan Hospital, Shanghai Public Health Clinical Center
5. Shanghai Jiao Tong University (China, Asia): Renji Hospital
6. Fujian Medical University (China, Asia)
7. Fujian Provincial Hospital (China, Asia)
8. Fuzhou University (China, Asia)
9. Beth Israel Deaconess Medical Center (Israel, Asia)
10. University of Utah (USA, North America): Utah School of Medicine
In consideration of patient and institutional privacy and confidentiality, along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed. Specifically, every dataset in each task is collected from a subset of the institutions previously mentioned.

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).

Task1 employs all data in Tasks 2-5, and the details of the characteristics of the subjects are as described below.

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.

(Update) A case refers to MRI with golden standard myocardial pathology from Task 2, MRI with golden standard liver structure from Task 3, CT/MRI with whole heart structure annotation from Task 4, or MRI with golden standard LA structure and scar annotations from Task5.

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

Task1 employs all data in Tasks 2-5, and the details of the training/validation/testing data setting for each cohort are as described below.
(Update) The whole challenge data sets includes 1250+ cases from 18 different centers in Britain, France, China, and U.S.A. Detailed information of centers is provided in the Section of Data Source(s), Task 1.

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

Assessing and exploring the transferability and generalizabilty of foundation models or pre-trained models on medical images in practical scenarios is of great research value. Hence, we not only evaluate the method on different imaging domains, but spare both in-domain and out-of-domain dataset for testing. Details of data setting rules for each task are described below.

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.

None.

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 data had been manually delineated by doctor or well-trained observers who were not aware of the methodology of this work.

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.

For the segmentation label, the annotators were instructed to segment the ORIs slice-by-slice, using the brush tool in the ITK-SNAP.

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.

Experienced physicists specialized in cardiac MRI or PhD students who major in biomedical image processing. ***(Update) Each task was conducted by different radiologists specialized in this field. For each task, PhD students first independently finished the annotation for each case and the average annotation was then checked and amended by a senior radiologist***

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.

The final gold standard segmentation was achieved by averaging several manual delineations using the shape-based average approach if necessary.

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.

None.

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.

Inter- and intra-annotator variability will be offered.

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)

Dice Similarity Coefficient (DSC); Average Surface Distance (ASD); Hausdorff Distance (HD).

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

For segmentation and quantification:
DSC is well-suited for assessing the overlap or agreement between the predicted segmentation and the ground

truth, making it particularly useful when the regions of interest in medical images may partially overlap.
Both DSC and HD provide quantitative measures of segmentation accuracy, allowing for objective evaluation of segmentation algorithms. DSC quantifies the degree of overlap, while HD measures the maximum distance between two sets, indicating the degree of boundary mismatch.
ASD specifically focuses on measuring the average distance between the surfaces of the predicted segmentation and the ground truth and is sensitive to errors near the boundaries of segmented regions.
For classification, AUC assesses the model's ability to distinguish between positive and negative instances across various threshold values. Accuracy is a straightforward metric that measures the proportion of correctly classified instances out of the total.

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.

(Update)
The average of metrics, i.e., Dice Similarity Coefficient (DSC); Average Surface Distance (ASD); Hausdorff Distance (HD) is utilized to range performance.

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

Only complete submissions are evaluated.

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

(Update)We will set different awards for each task, i.e., best performance and best paper. Specifically, the best performance award only takes the quantitative results of participants into account, which is computed by averaging multiple metrics described in each track, while the best paper will be awarded based on the average of paper scores evaluated by reviewers.

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 mean and standard deviation, as well as all percentiles if needed, will be calculated for each group of results. Moreover, T-test will be used to determine if there is a significant difference between results of two participants, which would be helpful to analyze the final ranking results. ***(Updata) Specifically, nonparametric test such as Mann-Whitney U Test, KS test, or Bayes Factor will be utilized to determine the significance of top performing methods for small data.***

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

Mean and variance are common descriptive statistics to measure average performance and variability of the models over the test data.

A t-test is a type of inferential statistical method used to determine if there is a significant difference between the data of two groups, which would be helpful for the final ranking.

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.

None.

TASK 2: Myocardial Pathology 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.

Myocardial infarction (MI) is a major cause of mortality and disability worldwide. Assessment of myocardial viability is essential in the diagnosis and treatment management of MI patients. This track addresses myocardial pathology segmentation (MyoPS) with multi-sequence CMRs. The objective is to categorize myocardial regions into healthy myocardium, scar, and edema from multi-sequence CMRs of MI patients. The challenge emphasizes overcoming the inclusion of multi-continent data, missing sequences for some centers, and misalignments in multi-sequence CMRs. The track seeks innovative solutions that address these challenges, leveraging real-world multi-sequence CMR data to improve MyoPS.

Keywords

List the primary keywords that characterize the task.

Myocardial pathology segmentation, Missing Data, Multi-Sequence, Misaligned data, Generalizability

ORGANIZATION

Organizers

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

Wangbin Ding, School of Imaging, Fujian Medical University, Fuzhou, China.
Sihan Wang, School of Data Science, Fudan University, Shanghai, China.
Freddy Odille, CIC-IT 1433, INSERM, Université de Lorraine and CHRU Nancy, Nancy, France. IADI U1254, INSERM and Université de Lorraine, Nancy, France.
Bailiang Chen, CIC-IT 1433, INSERM, Université de Lorraine and CHRU Nancy, Nancy, France. IADI U1254, INSERM and Université de Lorraine, Nancy, France.
Yingliang Ma, School of Computing Sciences, University of East Anglia, Norwich, U.K.
Lianming Wu, Renji Hospital, School of Medicine, Shanghai Jiao Tong UniversityShanghai, China.
Shan Yang, Department of Radiology, Zhongshan Hospital, Fudan University, Shanghai, China.
Yinyin Chen, Department of Radiology, Zhongshan Hospital, Fudan University, Shanghai, China.
Xiahai Zhuang, School of Data Science, Fudan University, Shanghai, China.

b) Provide information on the primary contact person.

Wangbin Ding: dingwangbin@fjmu.edu.cn, Sihan Wang: 21110980009@m.fudan.edu.cn

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

Same as above

Challenge venue and platform

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

None.

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

https://cmt3.research.microsoft.com/CARE_2024

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

https://zmiclab.github.io/projects/CARE_2024/

Participation policies

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

only automatic methods

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.

No additional data allowed.

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.

Same as above

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

Same as above.

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.

Same as above.

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).

Same as above.

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.

Same as above.

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.

Same as above.

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

Same as above.

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).

Same as above

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)

Same as above.

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.

Same as above.

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

Same as above.

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.

Same as above.

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

Same as above.

Task category(ies)

State the task category(ies)

Examples:

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

Segmentation, Registration, Quantification

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.

Acute/ chronic myocardial infarction patients who underwent single/multi-sequence cardiac MRI scans.

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

The challenge cohort is same as the target cohort.

Imaging modality(ies)

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

Cine MRI: Offers distinct visualization of cardiac motions and clear boundaries.
LGE MRI: Highlights infarcted areas for effective diagnosis.
***(Update) T2w MRI: Highlights edema areas for treatment planning. ***

Context information

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

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

No additional information.

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

None

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.

***(Update) LV region shown in different MRI sequences. ***

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.

(Update) Scars and edema of Myo.

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, Reliability, Robustness.

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).

Philips Achieva 1.5T, Siemens Avanto 1.5T, Siemens Aera 1.5T, Philips Ingenia 1.5T.

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

Dataset 2.A: Images were acquired with Siemens, General Electric, and Philips. The in-plane resolution of LGE scan was (0.78-1.76)×(0.78-1.76) mm, and the slice thickness was 2.5-20 mm.
Dataset 2.B: Images were acquired with Philips Achieva 1.5T. The in-plane resolution of LGE scan was (0.70-1.36)×(0.70-1.36) mm, and the slice thickness was 8-10 mm. The in-plane resolution of cine scan was (1.19-1.45)×(1.19-1.45) mm, and the slice thickness was 8-10 mm.
Dataset 2.C: Images were acquired with Philips Achieva 1.5T. The in-plane resolution of LGE scan was

(0.63-1.37)×(0.63-1.37) mm, and the slice thickness was 8-10 mm. The in-plane resolution of cine scan was (1.32-1.45)×(1.32-1.45) mm, and the slice thickness was 10 mm.
Dataset 2.D: Images were acquired with Siemens Avanto 1.5T. The in-plane resolution of LGE scan was (1.41-1.72)×(1.41-1.72) mm, and the slice thickness was 8 mm. The in-plane resolution of cine scan was (1.77-2.08)×(1.77-2.08) mm, and the slice thickness was 8-9 mm.
Dataset 2.E: Images were acquired with Philips Achieva 1.5T. The in-plane resolution of LGE scan was 0.75×0.75 mm, and the slice thickness was 8 mm. The in-plane resolution of T2w scan was 1.35×1.35 mm, and the slice thickness was 12-20 mm. The in-plane resolution of cine scan was 1.25×1.25 mm, and the slice thickness was 8-13 mm.
Dataset 2.F: Images were acquired with Siemens Aera 1.5T. The in-plane resolution of LGE and T2w scan was 1.33×1.33 mm, and the slice thickness was 10 mm. The in-plane resolution of cine scan was 1.77×1.77 mm, and the slice thickness was 10 mm.
Dataset 2.G: Images were acquired with Philips Ingenia 1.5T. The in-plane resolution of LGE and T2w scan was 1.33×1.33 mm, and the slice thickness was 10 mm. The in-plane resolution of cine scan was 1.77×1.77 mm, and the slice thickness was 10 mm.

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.

The data for task 2 was collected from seven centers listed above across two continents.
In consideration of patient and institutional privacy and confidentiality, along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed.

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).

Experienced physicists specialized in cardiac MRI.

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 refers to MRI(s) with structure and myocardial pathology annotations.

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

***(Update)
The data for task 2 was collected from 7 centers, namely 2.A/B/C/D/E/F/G. Information of data sources is provided

in the Section of Data Source(s), Task 1. Note that in consideration of patient and institutional privacy and confidentiality, particularly along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed here.
Training + validation: 280 patients {Centers: 2.A=180, 2.B=10 2.C=10, 2.D=10, 2.E=45, 2.F=25}
Testing: 75 patients {Centers: 2.F=25, 2.G=50}
***

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

It's important to have sufficient training data to train models effectively, especially for a challenging task like myocardial pathology segmentation. The choice of the proportion of training and test data is crucial for ensuring reasonable performance and accurate evaluation of the developed algorithms.
Besides, one unseen dataset is reserved for testing, which could help assess how well the developed models can perform on previously unencountered images. This is an important aspect of evaluating the robustness and applicability of the algorithms in real-world scenarios.

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.

None.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

None.

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.

Same as above.

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)

Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), Average Surface Distance (ASD);

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

Same as above.

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.

(Update)
The average of metrics, i.e., Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Average Surface Distance (ASD) is utilized to range performance.

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

Same as above

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

Same as above.

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.

Same as above.

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

Same as above.

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.

None.

TASK 3: Liver Fibrosis Quantification and Analysis

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.

Liver fibrosis, arising from chronic viral or metabolic liver conditions, presents a significant global health challenge. Precise liver segmentation (LiSeg) and fibrosis staging (LiFS) are essential for evaluating disease severity and facilitating accurate diagnoses.
This task focuses on achieving automatic liver segmentation and fibrosis staging from multi-phase and multi-center liver MRI scans. Automatic LiFS faces challenges including missing modalities for certain patients, misalignments in multi-phase MRIs, and how to better integrate multi-phase information for achieving more precise and generalizable liver fibrosis diagnoses. The challenges encountered in automatic LiSeg include limited ground truth across different modalities and domain shifts in multi-center data. Moreover, the extensive use of external data and pre-trained models are encouraged in this track to support liver segmentation.

Keywords

List the primary keywords that characterize the task.

Liver Fibrosis, Classification, Segmentation, Multi-Sequence, Missing Data, Misaligned data, Generalizability

ORGANIZATION

Organizers

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

Jiyao Liu, Institute of Science and Technology for Brain-Inspired Intelligence
, Fudan University, Shanghai, China.
Nannan Shi, Shanghai Public Health Clinical Center, Fudan University,Shanghai, China. Yuanye Liu, School of Data Science, Fudan University, Shanghai, China.
Zheyao Gao, School of Data Science, Fudan University, Shanghai, China.
Yuxin Li, School of Data Science, Fudan University, Shanghai, China.
Yuxin Shi, Shanghai Public Health Clinical Center, Fudan University,Shanghai, China. Fei Shan, Shanghai Public Health Clinical Center, Fudan University,Shanghai, China. Xiahai Zhuang, School of Data Science, Fudan University, Shanghai, China

b) Provide information on the primary contact person.

Jiyao Liu: jiyaoliu.fudan@gmail.com, Yuanye Liu: yuanyeliu@fudan.edu.cn

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

Same as above

Challenge venue and platform

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

None.

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

https://cmt3.research.microsoft.com/CARE_2024

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

https://zmiclab.github.io/projects/CARE_2024/

Participation policies

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

only automatic methods

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.

Publicly available data is allowed.

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.

Same as above

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

Same as above.

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.

Same as above.

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).

Same as above.

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.

Same as above.

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.

Same as above.

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

Same as above.

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).

Same as above

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)

Same as above.

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.

Same as above.

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

Same as above.

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.

Same as above.

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

Same as above.

Task category(ies)

State the task category(ies)

Examples:

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

Classification, Segmentation, Registration

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.

Patients diagnosed with liver fibrosis underwent multi-phase MRI scans.

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

The challenge cohort is same as the target cohort.

Imaging modality(ies)

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

T2WI: Reflects the water distribution and fibrous tissue.
DWI: Offers the diffusion characteristics of water molecules.
Gd-EOB-DTPA-enhanced dynamic MRIs: Includes non-contrast phase, arterial phase, venous phase, delay phase, and hepatobiliary phase, which can reflect vascular structure and perfusion status in the liver.

Context information

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

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

No additional information.

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

None

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.

Liver region shown in abdominal MRI, including T2-weighted imaging, diffusion-weighted imaging, and gadolinium ethoxybenzyl diethylenetriamine pentaacetic acid (Gd-EOB-DTPA)-enhanced multi-phase dynamic MRIs.

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.

Liver fibrosis staging including 4 types of fibrosis severity.
Liver structure.

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, Reliability, Robustness, Generalizability.

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).

Philips Ingenia3.0T, Philips Ingenia3.0T, Siemens Skyra 3.0T, Siemens Aera 1.5T.

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

The image acquisition protocol is as follows:
For all images: FOV=280-380mm, Flip angle=9-10, matrix=180-320, slice thickness=2-3mm. For Philips Ingenia3.0T - A centre, Philips Ingenia3.0T - B centre, Siemens Skyra 3.0T - C centre, Siemens Aera 1.5T - D centre: TR/TE=3.2-3.4/0, 3.4-3.6/0, 1.3-1.5/3.4-3.8 and 3.8-4.0/1.2-1.4ms, respectively.

Contrast-enhanced scans were performed based on the injection of GD-EOB-DTPA agent. The arterial phase was when the contrast agent entered the left ventricle, the portal phase was after 1min, the venous phase was after 90s, and the delay phase was after 3 min, and the hepatobiliary phase was after 20 min.

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.

The data for task 3 was collected from four centers listed above from one continent.
In consideration of patient and institutional privacy and confidentiality, along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed.

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).

Experienced physicists specialized in liver MRI.

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 refers to MRIs, golden standard liver structure, and staging ground truth.

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

***(Update)
The data for task 3 was collected from four centers, namely, 3.A/B/C/D. Information of data sources is provided in the Section of Data Source(s), Task 1. Note that in consideration of patient and institutional privacy and confidentiality, particularly along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed here.

Training set: {3.A=176, 3.B=70, 3.C=122, total=368}, where only 10 of each center have liver segmentation ground truth
Validation set: {3.A=10, 3.B=10, 3.C=10, total=30}
Test set: {3.A=47, 3.B=21, 3.C=35, 3.D=196, total=299} Total: {3.A=233, 3.B=101, 3.C=167, 3.D=196, total=697} ***

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

Considering the generalization of unseen centers, it is necessary to add centers that have not been seen in the training and validation sets in the test set. For the seen centers in the liver segmentation task, participants are encouraged to make better use of the unlabeled data, foundation model, and external datasets, thus providing few training samples with segmentation ground truth and large-range training data without segmentation ground truth.

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.

None.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

None.

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.

Same as above.

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)

Classification: Area under curve (AUC); Accuracy;
Segmentation: Dice Similarity Coefficient (DSC), Hausdorff Distance

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

Same as above.

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.

(Update)
The performance of classification and segmentation is ranged separately.
For classification, the average of metrics, i.e., Area under curve (AUC) and accuracy; is utilized to range performance.
For segmentation, the average of metrics, i.e., Dice Similarity Coefficient (DSC) and Hausdorff Distance is utilized to range performance.

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

Same as above

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

Same as above.

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.

Same as above.

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

Same as above.

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.

None.

TASK 4: Whole Heart 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.

Cardiovascular diseases (CVDs), identified by the WHO as the leading global cause of death, necessitate precise morphological and pathological quantification through segmentation of crucial cardiac structures from medical images. This task focuses on achieving whole heart segmentation (WHS), extracting individual substructures including the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium of LV (Myo), ascending aorta (AO), or the entire aorta, and the pulmonary artery (PA).
Automated WHS faces challenges including heart shape variability during the cardiac cycle, clinical artifacts like motion and poor contrast-to-noise ratio, as well as domain shifts in multi-center data and the distinct modalities of CT and MRI. This task serves to inspire innovative solutions in the realms of biomedical imaging and computer vision, striving to overcome these challenges and advance automated WHS for enhanced understanding and treatment of CVDs.

Keywords

List the primary keywords that characterize the task.

Whole Heart Segmentation, Multi-Modality, Generalizability

ORGANIZATION

Organizers

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

Xicheng Sheng, School of Data Science, Fudan University, Shanghai, China.
Yingliang Ma, School of Computing Sciences, University of East Anglia, Norwich, U.K.
Yiming Chen, School of Data Science, Fudan University, Shanghai, China.
Lei Li, School of Electronics & Computer Science, University of Southampton, Southampton, UK.
Guang Yang, Bioengineering Department and Imperial-X, Imperial College London, London, UK.
Xiahai Zhuang, School of Data Science, Fudan University, Shanghai, China.

b) Provide information on the primary contact person.

Xicheng Sheng: xcsheng22@m.fudan.edu.cn, Hangqi Zhou: 21110980018@m.fudan.edu.cn

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

Same as above

Challenge venue and platform

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

None.

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

https://cmt3.research.microsoft.com/CARE_2024

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

https://zmiclab.github.io/projects/CARE_2024/

Participation policies

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

only automatic methods

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.

No additional data allowed.

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.

Same as above

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

Same as above.

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.

Same as above.

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).

Same as above.

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.

Same as above.

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.

Same as above.

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

Same as above.

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).

Same as above

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)

Same as above.

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.

Same as above.

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

Same as above.

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.

Same as above.

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

Same as above.

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.

Patients with a wide range of cardiac diseases (myocardium infarction, atrial fibrillation, atrioventricular block, tricuspid regurgitation, aortic valve stenosis, Alagille syndrome, Williams syndrome, dilated cardiomyopathy, aortic coarctation, and Tetralogy of Fallot) with cardiac CT or MRI scan.

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

The challenge cohort is same as the target cohort.

Imaging modality(ies)

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

The imaging modalities applied in this challenge are CT and MRI.

Context information

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

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

No additional information.

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

None

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.

The whole heart shown in CT or 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.

Seven substructures of the heart, including the LV blood cavity, the RV blood cavity, the LA blood cavity, the RA blood cavity, the myocardium of the LV, the ascending aorta and the pulmonary artery.

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, Reliability, Robustness.

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).

Philips 64-slice CT scanner, SIEMENS SOMATOM_Force CT scanner, GE Revolution_CT scanner, Philips Achieva 1.5T, Siemens Avanto 1.5T.

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

For Dataset4.A, the cardiac CT/CTA data were obtained from 64-slice Philips CT scanners using a standard coronary CT angiography protocol. The in-plane resolution of the axial slices is 0.78 × 0.78 mm, and the average slice thickness is 1.60 mm.
Dataset4.B, the cardiac CT/CTA data were obtained from a dual-source SIEMENS CT scanner or a a high-end, single-source GE CT scanner using standard coronary CT angiography protocols. The original resolution is not

available. The images were reconstructed to 1 x 1 x 1 mm. For Dataset4.C and Dataset4.D, the cardiac MRI data were acquired with a 1.5T Philips scanner or Siemens Avanto 1.5T scanner. A navigator-gated 3D balanced steady state free precession (b-SSFP) sequence was used for free-breathing whole heart imaging.
The data were acquired at a resolution various from (1.6 x 1.6 x 2) to (2 x 2 x 3.2) mm.

Dataset4.E, the cardiac MRI data were acquired with Philips Achieva 1.5T scanner. At this centre, a range of cardiac MRI sequences related to Steady-State Free Precession (SSFP) were employed for diverse imaging needs. Typical sequences include a free-breathing high-resolution 3D Turbo Field Echo with a resolution various from (0.84 x 0.84 x 1.5) to (0.89 x 0.89 x 2) mm, Balanced Turbo Field Echo with Breath-Hold with a resolution vairious from (1.18 x 1.18 x 6) to (1.21 x 1.21 x 6) mm.

For Dataset4.F, the cardiac MRI data were acquired with Siemens Avanto 1.5T scanner. A navigator-gated 3D balanced steady state free precession (b-SSFP) sequence was used for free-breathing whole heart imaging. The data were acquired at a resolution various from (0.76 x 0.76 x 1.5) to (0.92 x 0.92 x 1.5) mm.

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.

The data for task 4 was collected from five centers listed above across two continents.
In consideration of patient and institutional privacy and confidentiality, along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed.

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).

Experienced physicists specialized in cardiac MRI.

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 refers to a CT/MRI image with whole heart structure annotation.

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

***(Update)
The data for task 4 was collected from five centers, namely 4.A/B/C/D/E. Information of data sources is provided in the Section of Data Source(s), Task 1. Note that in consideration of patient and institutional privacy and

confidentiality, particularly along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed here.
***
Training set + Validation set: 116 patients {Centers: 4.A = 20, 4.B = 44, 4.C = 6 , 4.D=14, 4.E = 32}
Test set: 96 patients {Centers: 4.A = 40, 4.C = 13, 4.D = 27, 4.E = 16}

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

We introduce diverse new data to enrich the open-access dataset released in MM-WHS challenge 2017. For previously available data, we keep the original split. For new CT data, we assign images to training and validation sets. For new MRI data, we assign images from one center to training and validation sets, and the other to the test set. This approach not only ensures consistency in benchmarking but also critically assesses the generalizability of WHS algorithms.

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.

None.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

None.

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.

Same as above.

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)

Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), Average Surface Distance (ASD).

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

Same as above.

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.

(Update)
The average of metrics, i.e., Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Average Surface Distance (ASD) is utilized to range performance.

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

Same as above

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

Same as above.

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.

Same as above.

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

Same as above.

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.

None.

TASK 5: Left Atrial and Scar Quantification & 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.

Atrial fibrillation (AF) is the most common arrhythmia observed in clinical practice, occurring in up to 1% of the population and rising fast with advancing age. Understanding the position and extent of left atrial (LA) scars is crucial for unraveling the pathophysiology and progression of AF. The task aims to achieve (semi-)automatic segmentation of the LA cavity and the quantification of LA scars from multi-center LGE MRI scans. Challenges encompass multicenter variability, and inherent difficulties stemming from various LA shapes, thin walls, surrounding enhanced regions, and intricate scar patterns in AF patients.

Keywords

List the primary keywords that characterize the task.

Atrial Fibrillation, Left Artrium, Scar segmentation, Quantification, Generalizability

ORGANIZATION

Organizers

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

Lei Li, School of Electronics & Computer Science, University of Southampton, Southampton, UK.
Xingtao Lin, Department of College of Physics and Information Engineering
Fuzhou University, Fuzhou, China.
Liqin Huang, College of Physics and Information Engineering, Fuzhou University, Fuzhou, China.
Xiahai Zhuang, School of Data Science, Fudan University, Shanghai, China.

b) Provide information on the primary contact person.

Lei Li: lilei.sky@outlook.com, Yiming Chen: 23210850002@m.fudan.edu.cn

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

Same as above

Challenge venue and platform

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

None.

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

https://cmt3.research.microsoft.com/CARE_2024

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

https://zmiclab.github.io/projects/CARE_2024/

Participation policies

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

only and (semi-) automatic methods

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.

No additional data allowed.

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.

Same as above

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

Same as above.

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.

Same as above.

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).

Same as above.

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.

Same as above.

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.

Same as above.

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

Same as above.

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).

Same as above

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)

Same as above.

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.

Same as above.

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

Same as above.

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.

Same as above.

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

Same as above.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection

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

Segmentation, Quantification

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.

Atrial fibrillation patients who underwent cardiac LGE MRI pre- or after-ablation.

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

The challenge cohort is same as the target cohort.

Imaging modality(ies)

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

The imaging technique applied in this challenge is LGE MRI, where healthy and scar tissues are differentiated by their altered wash-in and wash-out contrast agent kinetics. Therefore, scar tissue is seen as a region of enhanced or high signal intensity to distinguish from healthy tissue.

Context information

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

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

No additional information.

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

None

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.

LA cavity and scars shown in LGE MRI.

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.

LA cavity region and scar regions in the LA wall.

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, Reliability, Robustness.

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).

Siemens Avanto 1.5T, Siemens Vario 3T, Philips Acheiva 1.5T.

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

Dataset5.A: The clinical images were acquired with Siemens Avanto 1.5T or Vario 3T using free-breathing (FB) with navigator-gating. The spatial resolution of one 3D LGE-MRI scan was 1.25 x 1.25 x 2.5 mm. The patient underwented an MR examination prior to ablation or was 3-6 months after ablation.
Dataset5.B: Current clinical images were acquired with Philips Acheiva 1.5T using FB and navigator-gating with fat suppression. The spatial resolution of one 3D LGE-MRI scan was 1.3 x 1.3 x 4.0 mm. The patient underwented an MR examination prior to ablation or was 3-6 months after ablation.
Dataset5.C: The clinical images were also acquired with Philips Acheiva 1.5T using FB and navigator-gating with fat suppression. The spatial resolution of one 3D LGE-MRI scan was 1.4 x 1.4 x 1.4 mm. The patient underwented an MR examination prior to ablation or was 1 month after ablation.

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.

The data for task 5 was collected from three centers listed above across three continents.
In consideration of patient and institutional privacy and confidentiality, along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed.

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).

Experienced physicists specialized in cardiac MRI.

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 refers to an MRI with golden standard LA structure and scar annotations.

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

***(Update)
The data for task 5 was collected from three centers, 5.A/B/C. Information of data sources is provided in the Section of Data Source(s), Task 1. Note that in consideration of patient and institutional privacy and confidentiality, particularly along with the study purpose of model generalizability, detailed information regarding the source of specific sub-datasets for each task will not be disclosed here.
***
A center has expressed willingness to include more cases. We are currently in the process of evaluating and incorporating these additional cases into our dataset. We expect a total of approximately 250 cases for the task.

For LA scar quantification,
Training + validation: 60+ patients (TBA) {Centers: 5.A = 60, 5.B = TBA}
Testing: 34 patients{Centers: 5.A = 34}

For LA segmentation,
Training + validation: 130+ patients (TBA) {Centers: 5.A = 130, 5.B = TBA}
Testing: 64 patients{Centers: 5.A = 24, 5.B = 20, 5.C = 20}

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

1.The total number of cases is limited, because few patients underwent scanning of LGE MRI. More importantly, labeling the images manually is laborious and time-consuming.

2.The task of this challenge is arduous, so sufficient training data is required to ensure reasonable performance. Therefore, the proportion of training and test is chosen.
3.The inclusion of both pre-ablation and post-ablation LGE MRI for LA segmentation, while solely requesting scar quantification on post-ablation LGE MRI, aligns with clinical needs of ablation efficacy assessment and tecnical considerations for segmentation accuracy.
4.Unknow domain in the training data will be validated in the validation stage, to test the model generalizability.
5.In this task, we include more cases to enrich the open-access dataset released in LAScarQS2022. Addition of new data to the training set enables consistent benchmark for longitudinal analysis, allowing for a precise evaluation of how emerging technologies and more diverse training data influence the evolution and performance of LA scar quantification algorithms.

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.

None.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

Same as above.

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.

None.

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.

Same as above.

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)

Dice Similarity Coefficient (DSC); Average Surface Distance (ASD); Hausdorff Distance (HD).

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

Same as above.

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.

(Update)
The average of metrics, i.e., Dice Similarity Coefficient (DSC); Average Surface Distance (ASD); Hausdorff Distance (HD) is utilized to range performance.

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

Same as above

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

Same as above.

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.

Same as above.

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

Same as above.

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.

None.

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.

N/A

Further comments

Further comments from the organizers.

N/A