Towards the Automatic Segmentation, Modeling and Meshing of the Aortic Vessel Tree from Multicenter Acquisitions: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Towards the Automatic Segmentation, Modeling and Meshing of the Aortic Vessel Tree from Multicenter Acquisitions

Challenge acronym

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

SEG.A.

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.

The aorta is the main artery of the human body and, with its branch arteries, it forms the aortic vessel tree (AVT) [1] and supplies the whole body with blood. Aortic diseases, like aneurysms and dissections, can lead to severe consequences if left untreated. Their treatment with open surgery is of high risk and therefore delayed with constant monitoring and drug treatments whenever possible. Yet, this requires a regular screening of the vessels for disease development [2]. The standard image modality for clinical assessment is computed tomography angiography (CTA), which provides a detailed view of the AVT. Optimally, the whole AVT geometry is reconstructed and compared with the geometry of the subsequent CTA scans. Not only to detect changes related to the pathology, but also to detect peripheral changes, either resulting from the primary pathology or new comorbidities.

However, manual execution of this task requires a slice-by-slice contouring, which can require up to a whole day for the aortic vessel tree of one scan, making this task unfeasible in clinical practice. Furthermore, an accurate reconstruction can be used to analyze the blood flow and the outcome of endovascular surgeries by means of numerical simulations. For this, AI-supported automatic segmentation methods have shown to be a possible solution, which can potentially run in real time or in the background of the clinical routine. An open problem is the translation of these algorithms to 1) work in several clinical institutions, because of different scanning protocols, especially with regards to scanning device, radiation dose and contrast agent, which lead to varying Hounsfield values in the AVT, and 2) rely on a limited amount of labelled data given the long annotation time. This challenge comprehends one main task and two optional subtasks. In the main task, we target the problem of vessel tree segmentation before the diagnosis of an aortic pathology. We provide the challenge participants with a

training set of AVTs and corresponding manual segmentations from three institutions. Participants are expected to design algorithms for an automatic AVT segmentation based on this training set. All the proposed methods will be evaluated based on a hidden test set from a fourth institution using Dice Similarity Score (DSC) and Hausdorff Distance (HD). The evaluation will consider the the variance and the sensitivity of the evaluation metrics (DSC and HD) generated by different CTA variabilities, such as intensities, rotations, translations, noise and artificial motion artifacts. For this, the ranking will also consider the quantitative sensitivity indices [7] to define how the proposed method copes with large input variation. B) The reconstruction of the AVTs needs to be ideally artifact free for visualization and blood flow simulation. In the first optional subtask, the reconstructed AVT surface geometries will be qualitatively evaluated by clinical specialists. The qualitative evaluation will focus on the number of branching arteries and on the visual quality of the produced results. In the second subtask, we expect the AVT reconstruction to be in the form of a volume mesh. This will be quantitatively evaluated in terms of mesh validity for applications of computational fluid dynamics and ranked using the scaled Jacobian and the number of mesh elements.

Challenge keywords

List the primary keywords that characterize the challenge.

multicenter, segmentation, shape modeling, visualization, meshing, aorta, CTA, artificial intelligence

Year

The challenge will take place in ...

2023

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none

Duration

How long does the challenge take?

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.

The expected number of participants with a valid submission is 25±10. The estimate is based on the public rankings of recent MICCAI challenges.

We formulate a clinically relevant task as a technical problem (semantic segmentation), so that participants with a technical background do not need in-depth medical knowledge of this domain (aorta). Additionally, we target the problem of vessel tree visualization -- e.g., the presence of artifacts that hinder the visual evaluation -- and the problem of volume mesh generation for computational medicine applications. These three complementary tasks will potentially provide a broad interest to the community from image processing to geometry computing and biomedical simulation.

Publication and future plans

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

Every challenge participant should submit at least one paper describing the implemented algorithms. If the contribution justifies the submission of multiple papers, the teams can submit a maximum of one paper per task. These papers will be peer-reviewed and the accepted papers will be published in a Springer LNCS challenge proceedings. We furthermore plan a challenge summary paper in a journal.

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.

The leaderboard will be calculated by the organizers before the conference (after the challenge submission deadline). At the conference, we will present these final rankings and award certificates and prices to the winners. For this presentation, we will bring an own laptop, but could need a projector, loud speakers and microphones at the conference site. The winning participants will also have the chance to present their solution in a short talk (hybrid). All other participants with accepted papers will be highly encouraged to present their work in the form of poster presentations.

TASK: Main task - Aortic Vessel Tree Segmentation in CTA Acquisitions from Several Clinical Centers

SUMMARY

Keywords

List the primary keywords that characterize the task.

Aortic vessel tree, segmentation, CTA

ORGANIZATION

Organizers

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

Antonio Pepe, Institute of Computer Graphics and Vision, Graz University of Technology, Austria.
Gian Marco Melito, Institute of Mechanics, Graz University of Technology, Austria.
Yuan Jin, Research Center for Connected Healthcare Big Data, China, ZhejiangLab, Hangzhou, China.
Jan Egger, AI-guided Therapies, Institute for AI in Medicine (IKIM), University Hospital Essen, Germany; Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), Essen, Germany.
Fen-hua Zhao, Department of Radiology, Affiliated Dongyang Hospital of Wenzhou Medical University, China.
Heinrich Mächler, Department of Cardiac Surgery, University Hospital Graz, Austria.
Jens Kleesiek, Medical Machine Learning, Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Germany; Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), Essen, Germany; German Cancer Consortium (DKTK), Partner Site Essen, Essen, Germany.
Gerhard A. Holzapfel, Institute of Biomechanics, Graz University of Technology, Austria; Department of Structural Engineering, Norwegian University of Science and Technology (NTNU), Norway.
Alejandro F. Frangi, Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK; Biomedical Imaging Department, Leeds Institute for Cardiovascular and Metabolic Medicine (LICAMM), School of Medicine, University of Leeds, Leeds, UK.

b) Provide information on the primary contact person.

Antonio Pepe (antonio.pepe@tugraz.at)

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

The grand-challenge platform (grand-challenge.org) will be used to set up the challenge website (Type 2).

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

Website will be set up on the grand-challenge platform, after acceptance.

Participation policies

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

Fully automatic.

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

To ease reproducibility, non-public data is allowed if it published before the proclamation of the winning teams. If the data is not made public and the team still obtains outstanding performance, we will still award them with a special mention.

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

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

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

We will award best paper & highest ranking certificates to the winning teams. We will also look for industrial sponsorship for money related awards.

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.

Top 3 performing methods will be announced publicly. Submitted papers describing methods and performance results will be published as Springer LNCS challenge proceedings. Participating teams can choose to contribute to a joint journal article that summarizes the challenge results.

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

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

We expect to coordinate a Springer LNCS proceeding. Members that contribute to the development of the algorithms or the writing of papers qualify as authors. Participating teams should submit their papers to the organizing committee of the challenge for peer-review. Accepted papers will be included in the Springer challenge

proceeding. Rejected papers can be contributed to other venues.

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 container submission (type 2) on Grand Challenge.

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.

We will follow a three-phase submission scheme. Participants can share their temporary training results during phase 1. During phase 2 they can test their algorithms on only 2 hidden cases up to 5 times. During phase 3 they can submit the final version, which will be tested on all hidden cases. For more details, refer to the answers to the challenge support team.

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

The registration, release of training and test cases: Starts on April 15, 2023 and ends on July 14, 2023. (Phase 1)
Results & paper submission opens: Starts on June 15, 2023 and ends on July 14, 2023 (Phase 2)
Results & paper submission opens: Starts on July 15, 2023 and ends on August 15 (Phase 3)

Ethics approval

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

The training data is already published online and IRB approval was obtained. IRB approval for the test set has been also obtained.

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.

Code availability

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

We will release the evaluation codes on GitHub.

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

Participants are highly encouraged to make their codes public.

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.

We will also look for company sponsorship for money related awards.

The organizers have access to the ground truth of the test cases.

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

CAD, Diagnosis, Longitudinal study, Research, Surgery.

Additional points: image segmentation

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

The target cohort are patients who undergo CT angiography without a previous diagnosis of aortic disease.

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

The AVT dataset [3] was constructed based on full-body CTA scans, which were taken from the KiTS19 Grand Challenge, the RIDER Lung CT dataset and retrospectively collected scans from the Dongyang Hospital.

Imaging modality(ies)

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

The imaging technique applied in the challenge are Computed Tomography Angiography (CTA) scans from the clinical routine.

Context information

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

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

The raw CTA images as well as the segmented aortic vessel tree volumes will be provided.

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

Males and females each account for approximately half of the datasets.

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.

Aortic Vessel Trees shown in Computed Tomography Angiography (CTA) data. The scans are from different institutions and were acquired with different radiation doses and devices. The final evaluation will be performed on scans from an additional institution.

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

The algorithm target is the aortic vessel tree segmented from CTA scans.

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)).

Feasibility, Applicability, Sensitivity, Runtime, Robustness, Integration in workflow.

Additional points: 1) Produce a segmentation method with good performance in terms of DSC and HD.
2) Produce a method with low sensitivity to image variations in terms of intensity, rotation, translation, noise and blur. Sensitivity to variation is measured using Sobol' indices [7].

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).

The AVT collection resembles data from the KiTS19 challenge, the RIDER challenge and additional CTA scans from the Dongyang Hospital (CN). The CTA scans from KiTS19 were collected during routine care of patients who were treated by either partial or radical nephrectomy at the University of Minnesota Medical Center. Many of the CTA scans were acquired at referring institutions and are therefore heterogeneous in terms of scanner manufacturers and acquisition protocols. The scans from RIDER were also collected during routine care of patients who were screened for lung cancer. The radiation dose and the device varies among the images. For both cases, more details about the single CTA scans are available on the respective websites. The scans from the Dongyang Hospital (CN) were collected using the equipment Philips brilliance 64 slice CT, based on the scanning parameters: tube voltage 120 kV, tube current 250 MAS, collimator width 0.625mm, pitch 1. The thickness of reconstruction layer was 2.5mm, and the interval was 2.5mm. We targeted this strong variation between the scans to aim at institution-

independent automatic segmentation of the aortic vessel tree.

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

The CTA scans from KiTS19 and RIDER were acquired at referring institutions and are therefore heterogeneous in terms of scanner manufacturers and acquisition protocols. In all cases, the patient was supine during image collection, and the height-width origins thus lie to the patients' left anteriors. When there were multiple qualifying series for a particular case, that with the smaller slice thickness was chosen, but slice thicknesses range from 1mm to 5mm. The CTA scans from the Dongyang Hospital were all acquired with the same device. All scans have the same slice thickness of 2.5mm.

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.

Part of the datasets are from the training set of the 2019 Kidney and Kidney Tumor Segmentation Challenge (KiTS19). All patients who underwent nephrectomy for kidney tumors but have healthy aortas at the University of Minnesota Medical Center between 2010 and 2018 were candidates for inclusion in this database. Additional scans were selected from the RIDER collection of patients with lung cancer. Finally, Fen-hua Zhao provided further retrospective CTA scans of the clinical routine from the Dongyang Hospital of Wenzhou Medical University (China).

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).

The CTA scans are obtained by experienced radiologists during the clinical routine and the aortic vessel trees are segmented manually. All segmentations were supervised and approved by a senior radiologist. The manual segmentation of the hidden test was also supervised and validated by a lead cardiovascular surgeon.

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.

One case refers to one CTA scan and a corresponding manual segmentation of the aortic vessel tree. One case also refers to an independent patient.

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

We provide 56 training cases from three different medical centers. In addition, we have 5 hidden cases from a further medical center for evaluation.

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

The chosen cases for training and evaluation can reflect the Hounsfield Units variabilities in four different clinical centers. Furthermore, the cases provide a large variability in terms of gender and ethnicity.

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

The cases resemble CTA acquisitions from different manufacturers and different geographical locations (USA, Europe, China).

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.

The semi-automatic segmentations of the aortic vessel trees have been done by one annotator using 3D Slicer with the GrowCut algorithm and refined with manual user intervention.

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

The overall workflow starts with selecting and loading an aortic CTA (.nrrd file) into Slicer. Afterwards, noise has been removed as a pre-processing step for the segmentation. The gradient anisotropic diffusion filter has been choosen for its capabilities in edge-preservance. Next, local thresholding was performed. Therefore, a threshold range via masking has been manually specified. A minimum diameter of 3mm was used and GrowCut has been selected as segmentation algorithm. Manual refinement was performed where necessary.

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

The annotations have been performed by a biomedical student and checked 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.

Segmentation refinements were performed directly in 3D Slicer.

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.

We directly provide the raw CTA scans and segmentations using the NRRD file format.

Sources of error

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

Due to the scanning resolution, very small aortic branches may not be visible in the scan.

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

Sources of error might rely in inhomogeneity of the contrast agent.
The final test set is carefully selected to avoid cases with strong inhomogeneities.

ASSESSMENT METHODS

Metric(s)

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

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

1. Dice Similarity Coefficient (DSC)
2. Hausdorff Distance (HD)
3. First- and total-order Sobol' sensitivity indices

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

DSC and HD are common metrics used in segmentation tasks [1,2].
Sobol' sensitivity indices are commonly employed in understanding the influence of model parameters on output variation [7].

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.

Both metrics and the sensitivity analysis indeces are ranked indepentely. For HD and DSC we consider average, mode, and variance of their distribution. For the sensitivity analysis, we consider first and total order indeces. A weighted average of such rankings is considered for the final ranking of the submissions. A detailed description and the formulas are provided in the answers to Reviewer 1, question 6.

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

The results will be generated on the hidden set. Only the algorithm is necessary.

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

The ranking based on DSC and HD is used for other MICCAI challenges like AutoImplant 2020 and 2021, as well as StructSeg 2019.
We introduce the sensitivity analysis to additionally rank the capacity of the proposed method to handle input variabilities. We also introduce a penalty term based on the Fisher's moment coefficients of skewness to prefer solutions characterized by lower metrics variance/skewness.

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.

We will exclude participants who fail to submit a working copy of their algorithm. The final ranking will be based on the performance on the hidden test. We will rely on mean and variance of DSC and HD, Sobol' sensitivity indices to assess whether the top performing algorithms are significantly better than the other algorithms.

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

The mean value of DSC and HD produced by the algorithms are indicators of their overall performance. The standard deviation measures the performance stability of the algorithms. The Sobol' sensitivity indices provide an overview of how sensitive the method is to input variation. We expect low sensitivities to input variations for methods that work automatically between clinics.

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.

Subtask 1: Visualization of the Aortic Vessel Tree

Subtask 2: Volumetric Meshing of the Aortic Vessel Tree

TASK: Subtask 1 - Visualization of the Aortic Vessel Tree

SUMMARY

Keywords

List the primary keywords that characterize the task.

Aortic vessel tree, visualization, shape modeling, meshing, computational fluid dynamics

ORGANIZATION

Organizers

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

Gian Marco Melito, Institute of Mechanics, Graz University of Technology, Austria.
Antonio Pepe, Institute of Computer Graphics and Vision, Graz University of Technology, Austria.
Jianning Li, Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Germany.
Jan Egger, AI-guided Therapies, Institute for AI in Medicine (IKIM), University Hospital Essen, Germany; Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), Essen, Germany.
Heinrich Mächler, Department of Cardiac Surgery, University Hospital Graz, Austria.
Jens Kleesiek, Medical Machine Learning, Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Germany; Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), Essen, Germany; German Cancer Consortium (DKTK), Partner Site Essen, Essen, Germany.
Alejandro F. Frangi, Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK; Biomedical Imaging Department, Leeds Institute for Cardiovascular and Metabolic Medicine (LICAMM), School of Medicine, University of Leeds, Leeds, UK.
Dieter Schmalstieg, Institute of Computer Graphics and Vision, Graz University of Technology, Austria.
Gerhard A. Holzapfel, Institute of Biomechanics, Graz University of Technology, Austria; Department of Structural Engineering, Norwegian University of Science and Technology (NTNU), Norway.

b) Provide information on the primary contact person.

Gian Marco Melito (gmelito@tugraz.at)

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

The grand-challenge platform (grand-challenge.org) will be used to set up the challenge website (Type 2).

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

Website will be set up on the grand-challenge platform, after acceptance.

Participation policies

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

Fully automatic.

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

To ease reproducibility, non-public data is allowed if it published before the proclamation of the winning teams. If the data is not made public and the team still obtains outstanding performance, we will still award them with a special mention.

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

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

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

For each subtask, we will award best paper & highest ranking certificates to the winning teams. We will also look for industrial sponsorship for money related awards.

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.

Top 3 performing methods will be announced publicly. Submitted papers describing methods and performance results will be published as Springer LNCS challenge proceedings. Participating teams can choose to contribute to a joint journal article that summarizes the challenge results.

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

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

We expect to coordinate a Springer LNCS proceeding. Members that contribute to the development of the algorithms or the writing of papers qualify as authors. Participating teams should submit their papers to the organizing committee of the challenge for peer-review. Accepted papers will be included in the Springer challenge proceeding. Rejected papers can be contributed to other venues.

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 container submission (type 2) on Grand Challenge.

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.

We will follow a three-phase submission scheme. Participants can share their temporary training results during phase 1. During phase 2 they can test their algorithms on only 2 hidden cases up to 5 times. During phase 3 they can submit the final version, which will be tested on all hidden cases. For more details, refer to the answers to the challenge support team.

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

The registration, release of training and test cases: Starts on April 15, 2023 and ends on July 14, 2023. (Phase 1)
Results & paper submission opens: Starts on June 15, 2023 and ends on July 14, 2023 (Phase 2)
Results & paper submission opens: Starts on July 15, 2023 and ends on August 15 (Phase 3)

Ethics approval

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

The training data is already published online and IRB approval was obtained.

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.

Code availability

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

We will release the evaluation codes on GitHub.

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

Participants are highly encouraged to make their codes public.

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.

We will also look for company sponsorship for money related awards.

The organizers have access to the ground truth of the test cases.

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

Education, Research, Screening, Intervention planning.

Additional points: computational medicine

Task category(ies)

State the task category(ies).

Examples:

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

Modeling.

Cohorts

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

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

The target cohort are patients who undergo CT angiography and for whom an accurate visual analysis of the aortic tree or a blood flow simulation is needed.

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

The AVT dataset [3] was constructed based on full-body CTA scans, which were taken from the KiTS19 Grand Challenge, the RIDER Lung CT dataset and retrospectively collected scans from the Dongyang Hospital.

Imaging modality(ies)

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

The imaging technique applied in the challenge are Computed Tomography Angiography (CTA) scans from the clinical routine.

Context information

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

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

The raw CTA images as well as the segmented aortic vessel tree volumes will be provided.

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

Males and females each account for approximately half of the datasets.

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.

Aortic Vessel Trees shown in Computed Tomography Angiography (CTA) data. The scans are from different institutions and were acquired with different radiation doses and devices. The final evaluation will be performed on scans from an additional institution.

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

The algorithm target is the aortic vessel tree segmented from CTA scans.

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, Complexity, User satisfaction.

Additional points: Provide a 3D visualization that qualitatively conveys all meaningful features of the aortic vessel tree. The qualitative evaluation will take place in the form of a Likert-scale questionnaires and will be evaluated from clinical specialists from Essen (DE) and Graz (AT). The complete questionnaire will be available at the challenge opening.

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).

The AVT collection resembles data from the KiTS19 challenge, the RIDER challenge and additional CTA scans from the Dongyang Hospital (CN). The CTA scans from KiTS19 were collected during routine care of patients who were treated by either partial or radical nephrectomy at the University of Minnesota Medical Center. Many of the CTA scans were acquired at referring institutions and are therefore heterogeneous in terms of scanner manufacturers and acquisition protocols. The scans from RIDER were also collected during routine care of patients who were screened for lung cancer. The radiation dose and the device varies among the images. For both cases, more details about the single CTA scans are available on the respective websites. The scans from the Dongyang Hospital (CN)

were collected using the equipment Philips brilliance 64 slice CT, based on the scanning parameters: tube voltage 120 kV, tube current 250 MAS, collimator width 0.625mm, pitch 1. The thickness of reconstruction layer was 2.5mm, and the interval was 2.5mm. We targeted this strong variation between the scans to aim at population and resolution-independent visualization and meshing of the aortic vessel tree.

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

The CTA scans from KiTS19 and RIDER were acquired at referring institutions and are therefore heterogeneous in terms of scanner manufacturers and acquisition protocols. In all cases, the patient was supine during image collection, and the height-width origins thus lie to the patients' left anteriors. When there were multiple qualifying series for a particular case, that with the smaller slice thickness was chosen, but slice thicknesses range from 1mm to 5mm. The CTA scans from the Dongyang Hospital were all acquired with the same device. All scans have the same slice thickness of 2.5mm.

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.

Part of the datasets are from the training set of the 2019 Kidney and Kidney Tumor Segmentation Challenge (KiTS19). All patients who underwent nephrectomy for kidney tumors but have healthy aortas at the University of Minnesota Medical Center between 2010 and 2018 were candidates for inclusion in this database. Additional scans were selected from the RIDER collection of patients with lung cancer. Finally, Fen-hua Zhao provided further retrospective CTA scans of the clinical routine from the Dongyang Hospital of Wenzhou Medical University (China).

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).

The CTA scans are obtained by experienced radiologists during the clinical routine and the aortic vessel trees are segmented manually. All segmentations were supervised and approved by a senior radiologist. The manual segmentation of the hidden test was also supervised and validated by a lead cardiovascular surgeon.

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.

One case refers to one CTA scan and a corresponding manual segmentation of the aortic vessel tree. One case also refers to an independent patient.

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

We provide 56 training cases from three different medical centers. In addition, we have 5 hidden cases from a further medical center for evaluation.

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

The chosen cases for training and evaluation can reflect the Hounsfield Units variabilities in four different clinical centers. Furthermore, the cases provide a large variability in terms of gender and ethnicity.

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

The cases resemble CTA acquisitions from different manufacturers and different geographical locations (USA, Europe, China).

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.

The semi-automatic segmentations of the aortic vessel trees have been done by one annotator using 3D Slicer with the GrowCut algorithm and refined with manual user intervention.

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

The overall workflow starts with selecting and loading an aortic CTA (.nrrd file) into Slicer. Afterwards, noise has been removed as a pre-processing step for the segmentation. The gradient anisotropic diffusion filter has been choosen for its capabilities in edge-preservance. Next, local thresholding was performed. Therefore, a threshold range via masking has been manually specified. A minimum diameter of 3mm was used and GrowCut has been selected as segmentation algorithm. Manual refinement was performed where necessary.

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

The annotations have been performed by a biomedical student and checked 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.

Segmentation refinements were performed directly in 3D Slicer.

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.

We directly provide the raw CTA scans and segmentations using the NRRD file format.

Sources of error

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

Due to the scanning resolution, very small aortic branches may not be visible in the scan.

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

Discretization and segmentation errors can produce artifacts.
We will evaluate the visualization and meshing algorithms directly on the ground truth data to avoid any error propagation from Task 1.

ASSESSMENT METHODS

Metric(s)

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

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

1. Qualitative evaluation of the visualization using a Likert-scale questionnaire.
2a. Scaled Jacobian and number of invalid elements.
2b. Lowest number of elements in the volume mesh.

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

The qualitative evaluation will assess the presence or absence of artifacts in the visualization as well as the correct visualization of important features: perfusion, caliber, number of branches. For computational fluid dynamics simulations, the number of elements affects the duration of the simulation. The scaled Jacobian is a common metric to assess the validity of a mesh element for these applications.

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.

For subtask 1 (meshing), the ranking will be based on the formula provided in the answer to question 7 of Reviewer 1. The formula takes into account mode, variance and skewness of the scaled Jacobian distribution and the number of invalid elements. In case of equal scores, we rank higher solutions with a smaller number of total mesh elements.

For visualization, the ranking will be based on the overall Likert score for each submission. The reconstructed shape for each case is evaluated by two experienced experts based on the actual clinical usability (on a 5-point scale, 1 denotes unusable and 5 denotes flawless). The cumulative sum of the subjective scores is the final score for the
clinical evaluation.

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

The results will be generated on the hidden set. Only the algorithm is necessary.

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

We rely on metrics used in multiple visualization [9] and meshing works [10].
The qualitative evaluation will follow the settings of task 2 in the MICCAI AutoImplant 2021 challenge.

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.

We will exclude participants who fail to submit a working copy of their algorithm. There will be two separate rankings. The visualization ranking will rely on the cumulative score on a Likert-scale questionnaire presented to clinical experts.

For subtask 1 (meshing), the ranking will be based on the formula provided in the answer to question 7 of Reviewer 1. The formula takes into account mode, variance and skewness of the scaled Jacobian distribution and the number of invalid elements (elements with negative scaled Jacobian).

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

User studies in the form of questionnaires are a common way to qualitatively assess the perceived quality of a visualization.
The scaled Jacobian is a common metric to assess the validity of mesh elements for numerical simulations.

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.

TASK: Subtask 2 - Volumetric Meshing of the Aortic Vessel Tree

SUMMARY

Keywords

List the primary keywords that characterize the task.

Aortic vessel tree, visualization, shape modeling, meshing, computational fluid dynamics

ORGANIZATION

Organizers

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

Gian Marco Melito, Institute of Mechanics, Graz University of Technology, Austria.
Antonio Pepe, Institute of Computer Graphics and Vision, Graz University of Technology, Austria.
Jianning Li, Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Germany.
Jan Egger, AI-guided Therapies, Institute for AI in Medicine (IKIM), University Hospital Essen, Germany; Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), Essen, Germany.
Heinrich Mächler, Department of Cardiac Surgery, University Hospital Graz, Austria.
Jens Kleesiek, Medical Machine Learning, Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Germany; Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), Essen, Germany; German Cancer Consortium (DKTK), Partner Site Essen, Essen, Germany.
Alejandro F. Frangi, Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK; Biomedical Imaging Department, Leeds Institute for Cardiovascular and Metabolic Medicine (LICAMM), School of Medicine, University of Leeds, Leeds, UK.
Dieter Schmalstieg, Institute of Computer Graphics and Vision, Graz University of Technology, Austria.
Gerhard A. Holzapfel, Institute of Biomechanics, Graz University of Technology, Austria; Department of Structural Engineering, Norwegian University of Science and Technology (NTNU), Norway.

b) Provide information on the primary contact person.

Gian Marco Melito (gmelito@tugraz.at)

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

The grand-challenge platform (grand-challenge.org) will be used to set up the challenge website (Type 2).

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

Website will be set up on the grand-challenge platform, after acceptance.

Participation policies

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

Fully automatic.

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

To ease reproducibility, non-public data is allowed if it published before the proclamation of the winning teams. If the data is not made public and the team still obtains outstanding performance, we will still award them with a special mention.

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

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

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

For each subtask, we will award best paper & highest ranking certificates to the winning teams. We will also look for industrial sponsorship for money related awards.

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.

Top 3 performing methods will be announced publicly. Submitted papers describing methods and performance results will be published as Springer LNCS challenge proceedings. Participating teams can choose to contribute to a joint journal article that summarizes the challenge results.

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

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

We expect to coordinate a Springer LNCS proceeding. Members that contribute to the development of the algorithms or the writing of papers qualify as authors. Participating teams should submit their papers to the organizing committee of the challenge for peer-review. Accepted papers will be included in the Springer challenge proceeding. Rejected papers can be contributed to other venues.

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 container submission (type 2) on Grand Challenge.

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.

We will follow a three-phase submission scheme. Participants can share their temporary training results during phase 1. During phase 2 they can test their algorithms on only 2 hidden cases up to 5 times. During phase 3 they can submit the final version, which will be tested on all hidden cases. For more details, refer to the answers to the challenge support team.

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

The registration, release of training and test cases: Starts on April 15, 2023 and ends on July 14, 2023. (Phase 1)
Results & paper submission opens: Starts on June 15, 2023 and ends on July 14, 2023 (Phase 2)
Results & paper submission opens: Starts on July 15, 2023 and ends on August 15 (Phase 3)

Ethics approval

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

The training data is already published online and IRB approval was obtained.

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.

Code availability

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

We will release the evaluation codes on GitHub.

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

Participants are highly encouraged to make their codes public.

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.

We will also look for company sponsorship for money related awards.

The organizers have access to the ground truth of the test cases.

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

Education, Research, Screening, Intervention planning.

Additional points: computational medicine

Task category(ies)

State the task category(ies).

Examples:

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

Modeling.

Cohorts

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

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

The target cohort are patients who undergo CT angiography and for whom an accurate visual analysis of the aortic tree or a blood flow simulation is needed.

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

The AVT dataset [3] was constructed based on full-body CTA scans, which were taken from the KiTS19 Grand Challenge, the RIDER Lung CT dataset and retrospectively collected scans from the Dongyang Hospital.

Imaging modality(ies)

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

The imaging technique applied in the challenge are Computed Tomography Angiography (CTA) scans from the clinical routine.

Context information

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

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

The raw CTA images as well as the segmented aortic vessel tree volumes will be provided.

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

Males and females each account for approximately half of the datasets.

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.

Aortic Vessel Trees shown in Computed Tomography Angiography (CTA) data. The scans are from different institutions and were acquired with different radiation doses and devices. The final evaluation will be performed on scans from an additional institution.

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

The algorithm target is the aortic vessel tree segmented from CTA scans.

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, Complexity, User satisfaction.

Additional points: Provide a volumetric mesh representation of the aortic vessel tree for computational fluid dynamics. The best method will be selected based on the highest scaled Jacobian and lowest number of elements. The lowest number of elements will be considered in case of equal score with the scaled Jacobian.

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).

The AVT collection resembles data from the KiTS19 challenge, the RIDER challenge and additional CTA scans from the Dongyang Hospital (CN). The CTA scans from KiTS19 were collected during routine care of patients who were treated by either partial or radical nephrectomy at the University of Minnesota Medical Center. Many of the CTA scans were acquired at referring institutions and are therefore heterogeneous in terms of scanner manufacturers and acquisition protocols. The scans from RIDER were also collected during routine care of patients who were screened for lung cancer. The radiation dose and the device varies among the images. For both cases, more details about the single CTA scans are available on the respective websites. The scans from the Dongyang Hospital (CN) were collected using the equipment Philips brilliance 64 slice CT, based on the scanning parameters: tube voltage

120 kV, tube current 250 MAS, collimator width 0.625mm, pitch 1. The thickness of reconstruction layer was 2.5mm, and the interval was 2.5mm. We targeted this strong variation between the scans to aim at population and resolution-independent visualization and meshing of the aortic vessel tree.

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

The CTA scans from KiTS19 and RIDER were acquired at referring institutions and are therefore heterogeneous in terms of scanner manufacturers and acquisition protocols. In all cases, the patient was supine during image collection, and the height-width origins thus lie to the patients' left anteriors. When there were multiple qualifying series for a particular case, that with the smaller slice thickness was chosen, but slice thicknesses range from 1mm to 5mm. The CTA scans from the Dongyang Hospital were all acquired with the same device. All scans have the same slice thickness of 2.5mm.

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.

Part of the datasets are from the training set of the 2019 Kidney and Kidney Tumor Segmentation Challenge (KiTS19). All patients who underwent nephrectomy for kidney tumors but have healthy aortas at the University of Minnesota Medical Center between 2010 and 2018 were candidates for inclusion in this database. Additional scans were selected from the RIDER collection of patients with lung cancer. Finally, Fen-hua Zhao provided further retrospective CTA scans of the clinical routine from the Dongyang Hospital of Wenzhou Medical University (China).

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).

The CTA scans are obtained by experienced radiologists during the clinical routine and the aortic vessel trees are segmented manually. All segmentations were supervised and approved by a senior radiologist. The manual segmentation of the hidden test was also supervised and validated by a lead cardiovascular surgeon.

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.

One case refers to one CTA scan and a corresponding manual segmentation of the aortic vessel tree. One case also refers to an independent patient.

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

We provide 56 training cases from three different medical centers. In addition, we have 5 hidden cases from a

further medical center for evaluation.

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

The chosen cases for training and evaluation can reflect the Hounsfield Units variabilities in four different clinical centers. Furthermore, the cases provide a large variability in terms of gender and ethnicity.

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

The cases resemble CTA acquisitions from different manufacturers and different geographical locations (USA, Europe, China).

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.

The semi-automatic segmentations of the aortic vessel trees have been done by one annotator using 3D Slicer with the GrowCut algorithm and refined with manual user intervention.

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

The overall workflow starts with selecting and loading an aortic CTA (.nrrd file) into Slicer. Afterwards, noise has been removed as a pre-processing step for the segmentation. The gradient anisotropic diffusion filter has been choosen for its capabilities in edge-preservance. Next, local thresholding was performed. Therefore, a threshold range via masking has been manually specified. A minimum diameter of 3mm was used and GrowCut has been selected as segmentation algorithm. Manual refinement was performed where necessary.

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

The annotations have been performed by a biomedical student and checked 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.

Segmentation refinements were performed directly in 3D Slicer.

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.

We directly provide the raw CTA scans and segmentations using the NRRD file format.

Sources of error

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

Due to the scanning resolution, very small aortic branches may not be visible in the scan.

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

Discretization and segmentation errors can produce artifacts.
We will evaluate the visualization and meshing algorithms directly on the ground truth data to avoid any error propagation from Task 1.

ASSESSMENT METHODS

Metric(s)

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

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

1. Qualitative evaluation of the visualization using a Likert-scale questionnaire.
2a. Scaled Jacobian and number of invalid elements.
2b. Lowest number of elements in the volume mesh.

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

The qualitative evaluation will assess the presence or absence of artifacts in the visualization as well as the correct visualization of important features: perfusion, caliber, number of branches. For computational fluid dynamics simulations, the number of elements affects the duration of the simulation. The scaled Jacobian is a common metric to assess the validity of a mesh element for these applications.

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.

For subtask 1 (meshing), the ranking will be based on the formula provided in the answer to question 7 of Reviewer 1. The formula takes into account mode, variance and skewness of the scaled Jacobian distribution and the number of invalid elements. In case of equal scores, we rank higher solutions with a smaller number of total mesh elements.

For visualization, the ranking will be based on the overall Likert score for each submission. The reconstructed shape for each case is evaluated by two experienced experts based on the actual clinical usability (on a 5-point scale, 1 denotes unusable and 5 denotes flawless). The cumulative sum of the subjective scores is the final score for the
clinical evaluation.

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

The results will be generated on the hidden set. Only the algorithm is necessary.

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

We rely on metrics used in multiple visualization [9] and meshing works [10].
The qualitative evaluation will follow the settings of task 2 in the MICCAI AutoImplant 2021 challenge.

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.

We will exclude participants who fail to submit a working copy of their algorithm. There will be two separate rankings. The visualization ranking will rely on the cumulative score on a Likert-scale questionnaire presented to clinical experts.

For subtask 1 (meshing), the ranking will be based on the formula provided in the answer to question 7 of Reviewer 1. The formula takes into account mode, variance and skewness of the scaled Jacobian distribution and the number of invalid elements (elements with negative scaled Jacobian).

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

User studies in the form of questionnaires are a common way to qualitatively assess the perceived quality of a visualization.
The scaled Jacobian is a common metric to assess the validity of mesh elements for numerical simulations.

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.

ADDITIONAL POINTS

References

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

[1] Jin, Yuan, et al. "Ai-based aortic vessel tree segmentation for cardiovascular diseases treatment: status quo." arXiv preprint arXiv:2108.02998 (2021).
[2] Pepe, Antonio, et al. "Detection, segmentation, simulation and visualization of aortic dissections: A review." Medical image analysis 65 (2020): 101773.
[3] Radl, Lukas, et al. "AVT: Multicenter aortic vessel tree CTA dataset collection with ground truth segmentation

masks." Data in Brief 40 (2022): 107801.
[4] Maier-Hein, Lena, et al. "Why rankings of biomedical image analysis competitions should be interpreted with care." Nat. Commun. 9(1), (2018): 5217.
[5] Reinke, Annika, et al. "How to exploit weaknesses in biomedical challenge design and organization." Int. Conf. Med. Image Comput. Comput. Assis. Interv., Springer, Cham (2018).
[6] Maier-Hein, Lena, et al. "BIAS: Transparent reporting of biomedical image analysis challenges." Medical Image Analysis, (2020): 101796.
[7] Iooss, B., Lemaître, P.. A Review on Global Sensitivity Analysis Methods. In: Dellino, G., Meloni, C. (eds) Uncertainty Management in Simulation-Optimization of Complex Systems. Operations Research/Computer Science Interfaces Series, vol 59. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7547-8_5
[8] Saltelli, Andrea, et al. "Global sensitivity analysis: the primer." John Wiley & Sons, 2008.
[9] Oeltze, Steffen, and Bernhard Preim. "Visualization of vasculature with convolution surfaces: method, validation and evaluation." IEEE Transactions on Medical Imaging 24.4 (2005): 540-548.
[10] Tarjuelo-Gutierrez, Jaime, et al. "High-quality conforming hexahedral meshes of patient-specific abdominal aortic aneurysms including their intraluminal thrombi." Medical & biological engineering & computing 52.2 (2014): 159-168.