Cerebral Aneurysm Detection and Analysis: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Cerebral Aneurysm Detection and Analysis

Challenge acronym

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

CADA

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.

Cerebral aneurysms are local dilatations of arterial blood vessels caused by a weakness of the vessel wall.
Subarachnoid hemorrhage (SAH) caused by the rupture of a cerebral aneurysm is a life-threatening condition associated with high mortality and morbidity. The mortality rate is above 40%, and even in case of survival cognitive impairment can affect patients for a long time.
It is therefore highly desirable to detect aneurysms early and decide about the appropriate rupture prevention strategy. Diagnosis and treatment planning are based on angiographic imaging using MRI, CT, or X-ray rotation angiography.
Major goals in image analysis are the detection and risk assessment of aneurysms. We, therefore, subdivided the challenge into three categories. The first task is finding the aneurysm; the second task is the accurate segmentation to allow for a longitudinal assessment of the development of suspicious aneurysms. The third task is the estimation of the rupture risk of the aneurysm.

Challenge keywords

List the primary keywords that characterize the challenge.

aneurysm; automatic detection; segmentation; quantification; X-ray rotational angiography; stroke risk; risk prediction

Year

The challenge will take place in ...

2020

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.

We expect interest by participants from medical image processing as well as computational neuroscience and CFD. Considering the travel costs we expect about 25 to 30 participants.
We will invite participants from related events such as the Interdisciplinary Cerebrovascular Symposium explicitly.

Publication and future plans

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

We plan to organize proceedings with the papers corresponding to the submitted solutions in LNCS.
The challenge results will be published in two separate journal submissions. The first one will focus on the results of the detection and segmentation challenges. The second paper will present the results for the risk prediction.

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 results will be calculated offline before the challenge workshop takes place, so that we can focus on presentations and discussions. We therefore only need a projector or big monitor, microphones and loudspeakers.

TASK: Automatic Aneurysm Detection

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.

Cerebral aneurysms are local dilatation of arterial blood vessels, which can lead to life-threatening hemorrhage. It is therefore highly desirable to detect aneurysms early to enable preventive treatment. The goal of this task is to provide methods, which support clinicians in finding all aneurysms in a data set.

Keywords

List the primary keywords that characterize the task.

computer-aided detection; aneurysm; rotational X-ray angiography;

ORGANIZATION

Organizers

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

Tabea Kossen, Charité – Universitätsmedizin Berlin
Lilli Kaufhold, Charité – Universitätsmedizin Berlin
Markus Hüllebrand, Charité – Universitätsmedizin Berlin
Dr.-Ing. Jan-Martin Kuhnigk, Fraunhofer MEVIS
Jan Brühning, Charité – Universitätsmedizin Berlin
Jens Schaller, Charité – Universitätsmedizin Berlin
Boris Pfahringer, German Heart Center Berlin
Dr. med. Andreas Spuler, Helios Klinikum Berlin-Buch
Prof. Dr.-Ing. Leonid Goubergrits, Charité – Universitätsmedizin Berlin
Prof. Dr.-Ing. Anja Hennemuth, Charité – Universitätsmedizin Berlin

b) Provide information on the primary contact person.

Anja Hennemuth, anja.hennemuth@charite.de

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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.

grand-challenge.org

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

To appear.

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.

No policy defined.

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.

The two best submissions per submission type (docker container, result upload) with regard to the F2-score will be awarded.

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.

We intend to announce the five top-performing methods per class publicly at the workshop. All participants will receive their results as well as their ranking position per email and can decide whether they prefer an anonymous appearance of their method in the ranking table or if they do not want to make their result public at all.

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 intend to include two members per team as co-authors and expect the team to specify the contribution of these team members. We expect the contributors to wait with journal submissions on their individual approaches until the challenge papers have been accepted in order to be able to cite them.

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.

We will support two types of submissions:
a. The upload of (zipped) executable Docker containers fulfilling our API specification before the release of the test data
b. Upload of processed test data results in the specified formats.
The type of submission will be displayed in the result table.

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.

Given a sufficient number of submissions, rankings will be separate by submission type. In any case, the type of submission will be prominently indicated in the results. The evaluation will be performed with the final upload of each group. During the preparation phase participants can provide up to 5 test uploads, which will be checked for interface problems, data formats, etc.

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

• Challenge announcement and website with registration information: April 1
• Release of training data: April 14
• Platform open for test uploads (check of interfaces, data formats, ..., no feedback on performance will be provided): May 1
• Technical support for uploads until July 14
• Final submission of docker containers for automatic processing: July 5
• Release of test data: July 6
• Submission of automatic processing results: July 24
• Release of results: MICCAI Challenge Workshop 2020

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 data provided for this challenge is anonymized according to current regulations (removal of DICOM information as well as the facial part in the anatomical image data).
The ethics committee of Charité – Universitätsmedizin Berlin gave approval under ID EA2/222/19.

Data usage agreement

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

Examples:

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

CC BY NC ND.

Additional comments: Datasets will be made available to registered users under CC BY-NC-ND 4.0 license, meaning that data should only be shared between registered users, neither redistribution nor change or reuse is allowed. This license is only valid as long as regulations for data anonymization do not change.

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 code for calculating the validation metrics and their statistics is developed using a python and/or MeVisLab-backend in the COMIC environment.
A specification of the underlying calculations will be provided via the challenge web page.

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

Teams decide themselves whether they want to provide links to code repositories together with the description (paper) of their approach.

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 challenge is sponsored by the German Federal Ministry for Education and Research (Berlin Institute for the Foundations of Learning and Data (BIFOLD)) and the Fraunhofer Society.
The test labels will be available to registered participants according to the announced timeline described above.
The organizers listed in the proposal have access to all image data, segmentations, gender, age and rupture status per aneurysm. Other persons might have had access to subsets of the data in the following context:
- Image acquisition and clinical diagnosis: involved clinicians might be able to recognize datasets although they are completely anonymized, segmentation masks are not available in the PACS system
- CFD simulation using segmented surfaces: members of the biofluid mechanics lab at the ICM at Charité have tested methods with subsets of the dataset
- Training of segmentation methods with image data and label masks: students at TU Berlin will have temporary access to the provided dataset in the context of a teaching project. The schedule for releasing the data will follow the time plan of the challenge.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Assistance, Treatment planning, Screening, Decision support, CAD.

Task category(ies)

State the task category(ies).

Examples:

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

Detection.

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 humans who undergo imaging of the cranial arteries.
The goal is to find pathological dilatations of the vessels and enable their diagnostic assessment.

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

The shared dataset represents a cohort of 115 patients who underwent rotational X-ray angiography because of suspected SAH.

Imaging modality(ies)

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

Datasets have been acquired with a fixed C-Arm, which allows for the acquisition of CT-like 3D image volumes during diagnostic or interventional catheterization. Contrast agent was applied in the supplying artery of the examined vascular region (e.g. Vertebral artery). Image volumes show the arterial intracranial vasculature of the corresponding part of the brain.

Context information

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

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

In the training dataset, we store the positions of the aneurysm bounding box centers together with the image data. Furthermore, the segmented aneurysm heads are provided as stl- and mask-files.

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.

Originally, rotational X-ray angiography image data of the challenge cohort covered the patients' heads. The preprocessed (anonymized/cropped) datasets show the arterial subtree of the cranial vasculature. Patients underwent imaging because of suspected hemorrhage caused by the rupture of the vessel wall at a dilated vessel region (aneurysm) or for monitoring a known aneurysm.
For the target cohort, uncropped datasets might be processed directly.

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 algorithms are designed to detect image regions, which show pathological vessel dilatations (aneurysms).
For the aneurysm detection challenge, we expect participants to submit for each case per aneurysm a point coordinate representing the location of the aneurysm (ideally the center of the bounding box) in world coordinates. Additional two vectors describe the orientation and extent of the bounding box.

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)).

Hardware requirements, Sensitivity, Precision, Runtime.

Additional points: The goal is to find aneurysms and provide a bounding box, which contains the aneurysm with minimal margin. The ranking will be calculated based on the F2-score assessing recall and precision of the detection (provided point coordinate is inside aneurysm mask). The bounding box is assessed via the distance from the mask along the provided axes.

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 image data were acquired utilizing the digital subtraction AXIOM Artis C-arm system.

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

Data was acquired with a rotational acquisition time of 5 s with 126 frames (190° or 1.5° per frame, 1024 x 1024-pixel matrix, 126 frames). Post-processing was performed using LEONARDO InSpace 3D (Siemens, Forchheim, Germany). A contrast agent (Imeron 300, Bracco Imaging Deutschland GmbH, Germany) was manually injected into the internal carotid (anterior aneurysms) or vertebral (posterior aneurysms) artery. Reconstruction of a volume of interest selected by a neurosurgeon generated a stack of ~440 image slices with matrices of 512 x 512 voxels in-plane, resulting in an iso-voxel size of ~0.25 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.

Neurosurgery Department, Helios Klinikum Berlin-Buch.

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

Data was acquired by a neuroradiologist.

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 training case consists of an image dataset showing a contrast-enhanced cerebrovascular vessel tree.
Furthermore, segmentation masks (stl- and image files) are provided.
For test cases, only image datasets are provided.

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

23 cases will be used as test cases. The 92 cases released to the participants can be freely separated by the participants (e.g. for cross validation)

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

The total number of cases is the number of annotated datasets available at the time of submission of this challenge proposal. We intend to increase this number for later events. The number of test cases amounts to 20% of the available datasets.

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 probability of a point in the image domain representing a background voxel is much higher than for the aneurysm voxels. This has to be considered in the training design.

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.

An experienced annotator has provided manual segmentations. The windowing, orientation and zooming used for image annotation was not fixed. The users might have worked with different settings in the consecutive annotation sessions. The different settings might affect the assumed extensions of the aneurysms. An experienced neurosurgeon checked all segmentations.

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.

Because the original goal in segmentation aimed at using the surface meshes for CFD, the segmentation masks might be smoother than the actual vessel surface. Labels are available for all datasets.

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 initial annotation was performed by a biomedical engineer with 5 years of experience in the field of image-based modeling. He was supported by an experienced neurosurgeon.

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 second observer (clinical expert) corrected the result or the first observer if this was considered 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.

Data was anonymized (cropped) if necessary to achieve anonymization by face removal. Data has been converted into NIFTI format, so that all DICOM information is removed.

Sources of error

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

The second observer got the labeling result of the first observer as input, but was allowed to choose individual viewing settings. There might thus be a bias through the influence the first observer's result on the second observer. Furthermore, different data visualization settings might have limited the comparability of the visual detection of aneurysms by the observers.

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

The image data contains typical artefacts such as noise, ray artefacts, ... Furthermore, the coverage varies between datasets.

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)

For the assessment of the aneurysm detection we will calculate following metrics
-Recall R(true positive rate, sensitivity)
-Precision P(positive predictive value)
-Coverage C_cA of aneurysms cA by bounding boxes BB_cA
-Bounding box fit F_cA (max distance of bounding box from mask along main axes of the bounding box)

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

The major goal in detection is to make sure that aneurysms, which may pose a stroke risk, are not overlooked, so the sensitivity is an important measure. On the other hand, if the whole image is marked, the aneurysms would be included but the information would be meaningless, so the precision is important as well. A bounding box is helpful if it supports the visualization and postprocessing. To this end it should contain the aneurysm but be as small as possible.

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.

Ranking will be primarily based the F_2-score that combines recall R and precision P considering recall twice as important as precision:
F_2=5 PR/(4P+R)
Submissions with equal F2 score are further ranked according to the aneurysm coverage C_cA, and the bounding box fit F_cA is considered on the third level.
The authors are expected to perform cross-validation on the training dataset themselves.

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

During the submission test phase, submissions will be checked for completeness, and participants will be notified if cases are missing. During the validation phase missing cases will be interpreted as algorithm failure (worst possible metric value in each category).

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

The ranking score is chosen such that sensitivity is weighted stronger than precision, because missing a risk structure is considered worse than providing a false positive result. The coverage and bounding box fit are considered as second and third level measures for ranking of results with equal F_2-score because they are considered less important.

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.

During the submission test phase, submissions will be checked for completeness, and participants will be notified if cases are missing. During the validation phase missing cases will be interpreted as algorithm failure (worst possible metric value in each category).
The metrics per submission will be shown separately in a table with accompanying boxplots so that it is possible to compare the algorithm performance per class separately and analyze biases. For each metric, we will analyze the coefficient of variation.

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

No statistical method used for missing data

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.

We will analyze the achievable performance through combinations of submitted solutions. This might also help to assess the difficulty of this task.

TASK: Aneurysm 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.

The quantitative analysis for risk assessment and monitoring is usually based on the localization and segmentation of the aneurysm. Therefore, the goal of this task is to provide an accurate robust method for the segmentation of aneurysm heads.

Keywords

List the primary keywords that characterize the task.

segmentation; quantification; aneurysm; rotational X-ray angiography;

ORGANIZATION

Organizers

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

see Task 1

b) Provide information on the primary contact person.

Anja Hennemuth, anja.hennemuth@charite.de

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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.

grand-challenge.org

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

To appear.

Participation policies

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

Fully automatic, Semi 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.

No policy defined.

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.

The two best submissions per submission type (docker container, result upload) with regard to the combined normalized metrics will be awarded.

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.

see Task 1

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).

see Task 1

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.

see Task 1

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.

see Task 1

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

• Challenge announcement and website with registration information: April 1
• Release of training data: April 14
• Platform open for test uploads (check of interfaces, data formats, ..., no feedback on performance will be provided): May 1
• Technical support for uploads until July 14
• Final submission of docker containers for automatic processing: July 5
• Release of test data: July 6
• Submission of automatic processing results: July 24
• Release of initialization markers: July 25
• Submission of interactive processing results: July 30
• Release of results: MICCAI Challenge Workshop 2020

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).

see Task 1

Data usage agreement

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

Examples:

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

CC BY NC ND.

Additional comments: see Task 1

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.

see Task 1

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

see Task 1

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.

see Task 1

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, Intervention planning, Treatment planning, Diagnosis.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

The target cohort are humans who undergo imaging because of suspected subarachnoid hemorrhage (SAH) caused by ruptured aneurysms or known aneurysms, which are monitored. The quantitative assessment or aneurysm size and shape can be the basis for decision support in these patients.

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

see Task 1

Imaging modality(ies)

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

see Task 1

Context information

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

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

see Task 1

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.

see Task 1

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 goal is the delineation of the pathologically dilated part of the vessel (aneurysm).
The participants are expected to provide a label image with different labels for the segmented aneurysms. The coverage and voxel extents should match the original image data. For methods, which provide segmentations with subvoxel accuracy, results may also be stored in stl-files as closed surfaces.

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)).

Hardware requirements, Reliability, Accuracy.

Additional points: The segmentation should provide results, which reliably enable a quantitative assessment with regard to shape and volume of the aneurysm. The ranking is therefore based on a combined score that considers overlap, surface distance, volume correlation and volume bias of the results.

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).

see Task 1

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

see Task 1

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

see Task 1

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

see Task 1

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 training case consists of an image dataset showing a contrast-enhanced cerebrovascular vessel tree.
Furthermore, segmentation masks (stl- and image files) are provided.
For test cases, only image datasets and aneurysm locations (centerpoint coordinates) are provided (release after submission deadline of Task 1)

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

see Task 1

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

see Task 1

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 average volume per aneurysm is 0.391 ml, meaning that for image volumes between 280 and 2350 ml, there is a strong imbalance between foreground and background, which the participants have to consider when designing the preprocessing and training setup.

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.

see Task 1

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.

see Task 1

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.

see Task 1

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.

see Task 1

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.

see Task 1

Sources of error

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

The bias of the two-stage annotation as well as the lack of standardization of the visualization settings (see Task 1) during the annotation process might affect the delineation of the aneurysm surface as well. Annotations are represented as surfaces. The voxelization might induce further inexactness.

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

see 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)

For the assessment of the segmentation quality, we will compare the submission segmentation result masks M_cA^^* with ground truth masks M_cA from the expert annotations:
-Jaccard: J(M_cA^^*,M_cA)
-Hausdorff distance: HD(M_cA^^*,M_cA)
-Average distance: AVD(M_cA^^*,M_cA)
-Pearson correlation coefficient r between predicted V^^*and reference volume V of all aneurysms
-Bias (b) computed as the mean absolute difference of predicted and reference volume
-Standard deviation () of the difference between predicted and reference volumes.

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

The segmentation is the basis for the quantitative assessment of the aneurysms. It should enable the extraction of shape and volume parameters for the assessment of change over time or the comparison with decision thresholds. Therefore, the overlap, and distance from reference segmentations is important. For the assessment of volumes, we also analyze, how well the results correlate over the cohort and if there is a bias.

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 the ranking, we will perform a normalization according to the maximum among all participants so that each individual metric takes a value between 0 (worst case among all participants) and 1 (perfect fit between the reference and predicted segmentation). The ranking score is calculated as the average of the normalized metrics.

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

see Task 1

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

We consider the metrics described in 26a) as equally important for the application context and therefore try to integrate them with equal weight in the scoring system.

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.

see Task 1. We will further analyze, which or the equally weighted measures contributed most in the ranking process.

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

No statistical method used for missing data

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.

see Task 1

TASK: Rupture Risk Prediction

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.

The treatment strategy depends on the risk assessment of an aneurysm. This task aims at an image-based assessment of the aneurysms' rupture risk.

Keywords

List the primary keywords that characterize the task.

risk assessment; decision support; aneurysm; rotational X-ray angiography;

ORGANIZATION

Organizers

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

see Task 1

b) Provide information on the primary contact person.

Leonid Goubergrits, leonid.goubergrits@charite.de

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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.

grand-challenge.org

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

to appear

Participation policies

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

Fully interactive, Fully automatic, Semi 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.

No policy defined.

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.

The two best submissions per submission type (docker container, result upload) with regard to the F2-score will be awarded.

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.

see Task 1

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).

see Task 1

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.

see Task 1

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.

see Task 1

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

• Challenge announcement and website with registration information: April 1
• Release of training data: April 14
• Platform open for test uploads (check of interfaces, data formats, ..., no feedback on performance will be provided): May 1
• Technical support for uploads until July 14
• Submission of docker containers: July 30
• Release of test data: August 1
• Final submission of results: August 31
• Release of results: MICCAI Challenge Workshop 2020

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).

see Task 1

Data usage agreement

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

Examples:

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

CC BY NC ND.

Additional comments: see Task 1

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.

see Task 1

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

see Task 1

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.

see Task 1

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

Treatment planning, Decision support, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Prediction.

Cohorts

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

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

The target cohort are patients with known aneurysms. Risk prediction may help to decide whether immediate treatment is required.

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

see Task 1

Imaging modality(ies)

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

see Task 1

Context information

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

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

see Task 1

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

A table containing the information whether an aneurysm is ruptured is provided.

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.

see Task 1

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 algorithms will provide a risk classification for the dilated vessel wall based on the features shown in the image.
The participants provide a table listing the aneurysm position (provided as input) and the predicted rupture risk.

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)).

Hardware requirements, Sensitivity, Precision.

Additional points: The ranking will be calculated based on the F2-score (see Task 1).

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).

see Task 1

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

see Task 1

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

see Task 1

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

see Task 1

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 table with the aneurysm rupture information is provided (case, position, rupture state) in addition to the case data.
For the test cases, image data is released at the same time for Task 1 and 3, centerpoint coordinates with Task 2, and masks after the submission deadline for Task 2.

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

see Task 1

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

see Task 1

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 rupture information is stored in a table listing the aneurysms together with their rupture state (58% not ruptured vs 42% ruptured).

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 rupture information was provided by the treating neurosurgeon.

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.

None.

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 information about the rupture state was provided by the neurosurgeon.

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.

None.

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.

see Task 1

Sources of error

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

The information about the rupture status of the aneurysms is derived from text entries in the hospital information system. Errors in database queries and text interpretation could have introduced errors here. Furthermore, it is not clear whether aneurysms ruptured after the annotation and anonymization took place.

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

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)

For the assessment of the rupture risk prediction, we will calculate recall and precision with regard to the risk classification.

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

Same as in Task 1 an aneurysm at risk should not be misclassified, but the number of false alarms should also be low. Thus, sensitivity and precision are assessed.

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.

Ranking will be based on the F_2-score.

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

see Task 1

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

Same as in Task 1 the rupture risk of an aneurysm should not be overlooked. On the other hand, too many false positives mean a tedious screening for the physician, who has to review the risk assessment for decision making. The F_2-score combines recall and precision such that the identification of aneurysms at risk is considered more important than the avoidance of a false positive risk classification.

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.

see Task 1

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

No statistical method used for missing data

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 3 can use the results of 1 and 2 and might work best when combining different approaches.

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.

None

CADA – Cerebral Aneurysm Detection and Analysis

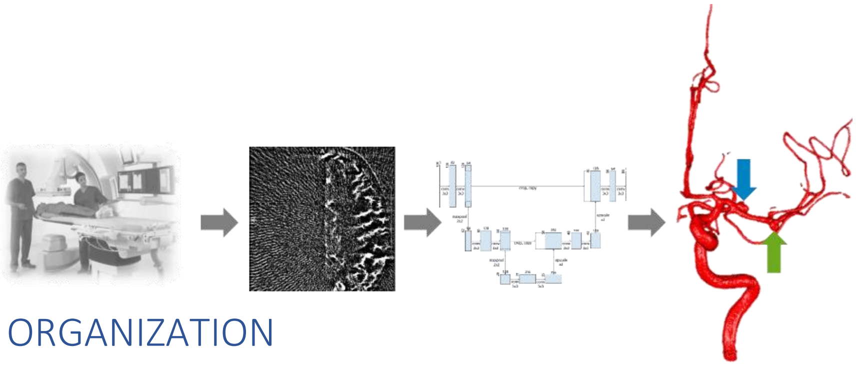
Abstract

Cerebral aneurysms are local dilations of arterial blood vessels caused by a weakness of the vessel wall. Subarachnoid hemorrhage (SAH) caused by the rupture of a cerebral aneurysm is a life-threatening condition associated with high mortality and morbidity. The mortality rate is above 40%, and even in case of survival cognitive impairment can affect patients for a long time.

It is therefore highly desirable to detect aneurysms early and decide about the appropriate rupture prevention strategy. Diagnosis and treatment planning are based on angiographic imaging using MRI, CT, or X-ray rotation angiography.

Major goals in image analysis are the detection and risk assessment of aneurysms. We, therefore, subdivided the challenge into three categories. The first task is finding the aneurysm; the second task is the accurate segmentation to allow for a longitudinal assessment of the development of suspicious aneurysms. The third task is the estimation of the rupture risk of the aneurysm.

Keywords: Aneurysm; automatic detection; quantification; X-ray rotational angiography; stroke risk; risk prediction



CHALLENGE ORGANIZATION

Team

Tabea Kossen, Charité – Universitätsmedizin Berlin
Lilli Kaufhold, Charité – Universitätsmedizin Berlin
Markus Hüllebrand, Charité – Universitätsmedizin Berlin
Dr.-Ing. Jan-Martin Kuhnigk, Fraunhofer MEVIS
Jan Brühning, Charité – Universitätsmedizin Berlin
Jens Schaller, Charité – Universitätsmedizin Berlin

Boris Pfahringer, German Heart Center Berlin

Dr. med. Andreas Spuler, Helios Klinikum Berlin-Buch

Prof. Dr.-Ing. Leonid Goubergrits, Charité – Universitätsmedizin Berlin
Prof. Dr.-Ing. Anja Hennemuth, Charité – Universitätsmedizin Berlin

Contact: Anja Hennemuth, anja.hennemuth@charite.de, Tel. +49 30 4593 2350, Leonid Goubergrits, leonid.goubergrits@charite.de, Tel: +49 30 450 553808

Lifecycle Type

This challenge will be initially organized as a one-time event with a fixed submission deadline at MICCAI 2020. Our envisioned goal is to extend the dataset with additional cases and modalities and potentially establish a recurring workshop event to support progress in this application field.

Challenge Venue and Platform

We intend to organize the challenge such that it is connected with a half day MICCAI workshop. For challenge management, we will use the Grand-challenge.org platform. In case of acceptance, we will provide a website for this challenge.

Participation Policies

The degree of user interactions allowed for the algorithms is chosen differently for the three application classes:

1. **Aneurysm Detection**
   No user interaction allowed. We will support submissions of docker containers (submission until July 5) with automatic methods as well as processing results stored in the specified format (submission until July 24).
2. **Aneurysm Segmentation**
   Semi-automatic methods are allowed. The degree of interaction, as well as the training of the persons involved in creating the results, must be described in the submission paper.
   Three types of submissions are accepted:
   - docker containers with automatic methods (submission until July 5)
   - automatic processing results (submission until July 24)
   - interactive processing results (submission until July 30)
3. **Rupture Risk Estimation**
   Semi-automatic methods are allowed. The degree of interaction, as well as the training of the persons involved in creating the results, must be described in the submission paper.
   types of submissions are accepted:
   - docker containers with automatic methods (submission until July 30)
   - processing results (submission until August 31)

Our dataset is relatively small, so it will be worthwhile to see how pretraining using different datasets might help. We encourage participants to make use of existing data and solutions. The use of such additional data for training or validation must be disclosed in the submission paper.

Members of the organization team and group members with an unfair advantage through earlier data access are allowed to submit their results but are not eligible for the awards.
We will award three prizes per class.

We intend to announce the five top-performing methods per class publicly at the workshop. All participants will receive their results as well as their ranking position per email and can decide whether they prefer a named or anonymous appearance of their method in the ranking table or if they do not want to make their result public at all.

We plan to organize proceedings with the papers corresponding to the submitted solutions in LNCS. The challenge results will be published in two separate journal submissions. The first one will focus on the results of the detection and segmentation challenges. The second paper will present the results for the risk prediction.

For the challenge papers, we intend to include up to two members per team as co-authors and expect the team to specify the contribution of these team members. We expect the contributors to wait with journal submissions on their individual approaches until the challenge papers have been accepted in order to be able to cite them.

Submission and Timeline

Timeline:

- Challenge announcement and website with registration information: April 1
- Release of training data: April 14
- Release of results: MICCAI Workshop 2020

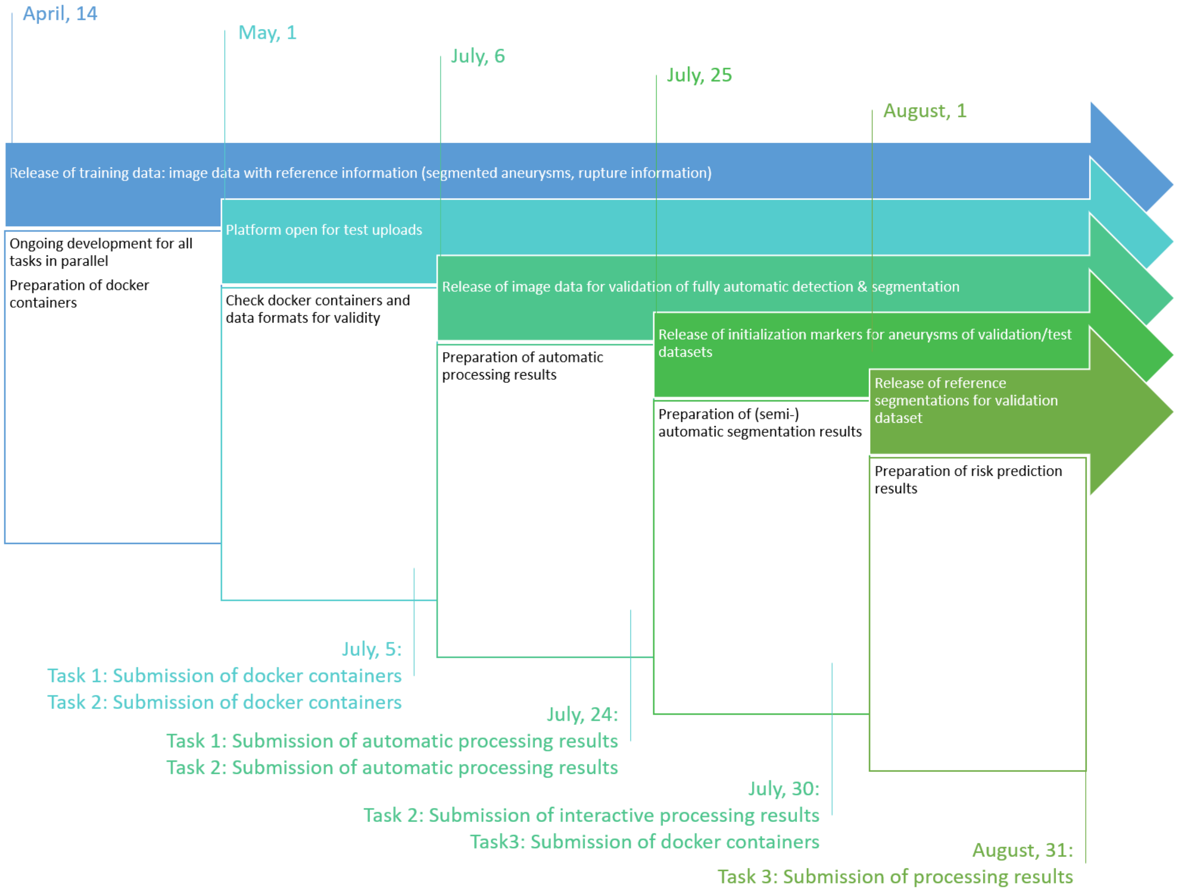

April, 14
May, 1
July, 6
July, 25
August, 1
Release of training data: image data with reference information (segmented aneurysms, rupture information)
Platform open for test uploads
Release of image data for validation of fully automatic detection & segmentation
Release of initialization markers for aneurysms of validation/test datasets
Release of reference segmentations for validation dataset
Ongoing development for all tasks in parallel
Preparation of docker containers
Check docker containers and data formats for validity
Preparation of automatic processing results
Preparation of (semi-) automatic segmentation results
Preparation of risk prediction results
July, 5:
Task 1: Submission of docker containers
Task 2: Submission of docker containers
July, 24:
Task 1: Submission of automatic processing results
Task 2: Submission of automatic processing results
July, 30:
Task 2: Submission of interactive processing results
Task3: Submission of docker containers
August, 31:
Task 3: Submission of processing results

Figure 1: Timeline of the challenge. Because we want to allow different types of approaches and submissions (docker containers and processed results), we have organized data release and submission in phases

We will accept submissions in two formats:

1. For automatic solutions, there are two options:
    a. The upload of (zipped) executable Docker containers fulfilling our API specification before the release of the test data.
    b. Upload of processed test data results in the specified formats.
2. For semi-automatic approaches, participants shall upload their processed test data results in the specified formats.

The evaluation will be performed with the final upload of each group. Given a sufficient number of submissions, rankings will be separate by submission type. In any case, the type of submission will be prominently indicated in the results.

Data

The data provided for this challenge is anonymized according to current regulations (removal of DICOM information as well as the facial part in the anatomical image data).

We received approval from the ethics committee of Charité – Universitätsmedizin Berlin under ID EA2/222/19 .

Datasets will be made available to registered users under CC BY-NC-ND 4.0 license, meaning that data should only be shared between registered users, neither redistribution nor change or reuse is allowed. This license is only valid as long as regulations for data anonymization do not change.

Code Availability

The code for calculating the validation metrics and their statistics is developed using a python and/or MeVisLab-backend in the COMIC environment.

A specification of the underlying calculations will be provided via the challenge web page.

Teams decide themselves whether they want to provide links to code repositories together with the description (paper) of their approach.

Conflicts of interest

The challenge is sponsored by the German Federal Ministry for Education and Research (Berlin Institute for the Foundations of Learning and Data (BIFOLD)) and the Fraunhofer Society.

The test labels will be available to registered participants according to the announced timeline described above.

MISSION OF THE CHALLENGE

The challenge addresses three major aspects of the inspection of angiographic cranial images, namely

- Detection of pathological changes of the vessel tree in the form of aneurysms
- The quantitative assessment of these aneurysms through segmentation as the basis for diagnosis and monitoring
- Image-based estimation of the stroke risk

These solutions could improve clinical workflows towards aneurysm screening in cranial angiographies, which might be acquired for a wide range of diagnostic questions.

Furthermore, they could improve the analysis of datasets acquired for patients with suspected subarachnoid hemorrhage (SAH) or known aneurysms, which are monitored.

The shared dataset represents a cohort of 115 patients who underwent rotational X-ray angiography because of suspected SAH.

Datasets have been acquired with a fixed C-Arm, which allows for the acquisition of CT-like 3D image volumes during diagnostic or interventional catheterization. Contrast agent was applied in the supplying artery of the examined vascular region (e.g. Vertebral artery). Image volumes show the arterial intracranial vasculature of the corresponding part of the brain.

In the training dataset, we store the positions of the aneurysm centers together with the image data. Furthermore, the segmented aneurysm heads are provided as stl- and mask-files. For each aneurysm the rupture state is provided.

In correspondence to the aspects introduced above, we will have three analysis goals:

1. Task 1: Automatically find position and approximate bounding box for all pathological vessel dilations (aneurysms) in a dataset to ensure that these structures, which are associated with stroke risk, are not missed during the inspection of a dataset
2. Task 2: Provide accurate segmentation masks for aneurysms (automatically or semi-automatically) in order to support the quantitative assessment for diagnosis, monitoring and therapy planning
3. Task 3: Classify aneurysms according to their rupture risk (using machine learning, computational geometry, CFD, …) to support decision making for treatment planning

CHALLENGE DATASETS

Data of 115 patients with cerebral aneurysms without vasospasm were collected for diagnostic and treatment decision purposes in the Neurosurgery Department of the Helios Hospital Berlin Buch.

The image data were acquired utilizing the digital subtraction AXIOM Artis C-arm system using a rotational acquisition time of 5 s with 126 frames (190° or 1.5° per frame, 1024 x 1024-pixel matrix, 126 frames). Post-processing was performed using LEONARDO InSpace 3D (Siemens, Forchheim, Germany). A contrast agent (Imeron 300, Bracco Imaging Deutschland GmbH, Germany) was manually injected into the internal carotid (anterior aneurysms) or vertebral (posterior aneurysms) artery. Reconstruction of a volume of interest selected by a neurosurgeon generated a stack of ~440 image slices with matrices of 512 x 512 voxels in-plane, resulting in an iso-voxel size of ~0.25 mm.

An experienced annotator has provided the segmentations. The windowing used for image annotation was not fixed. The users might have worked with different settings in the consecutive annotation sessions. The different settings might affect the assumed extensions of the aneurysms. An experienced neurosurgeon checked all segmentations. Because the original goal in segmentation aimed at using the surface meshes for CFD, the segmentation masks might be smoother than the actual vessel surface. Labels are available for all datasets.

Criterium for ruptured aneurysms is subarachnoid hemorrhage (SAH) seen on CT scans (most often) or MRI scans. In cases of normal CT or MRI scans but with symptoms suspicious of a subarachnoid hemorrhage, typical findings in cerebrospinal fluid (lumbar puncture) prove a subarachnoid hemorrhage. In cases of multiple aneurysms, the localization (asymmetry) of the SAH points to the ruptured aneurysm. If there is no asymmetry of the SAH, the more proximal or greater or, the more irregularly shaped aneurysm is considered to be the ruptured one. In rare cases, a xanthochrom parenchymal „halo“ around the aneurysm seen in surgery proves the history of a (small and past) hemorrhage originating from this aneurysm.

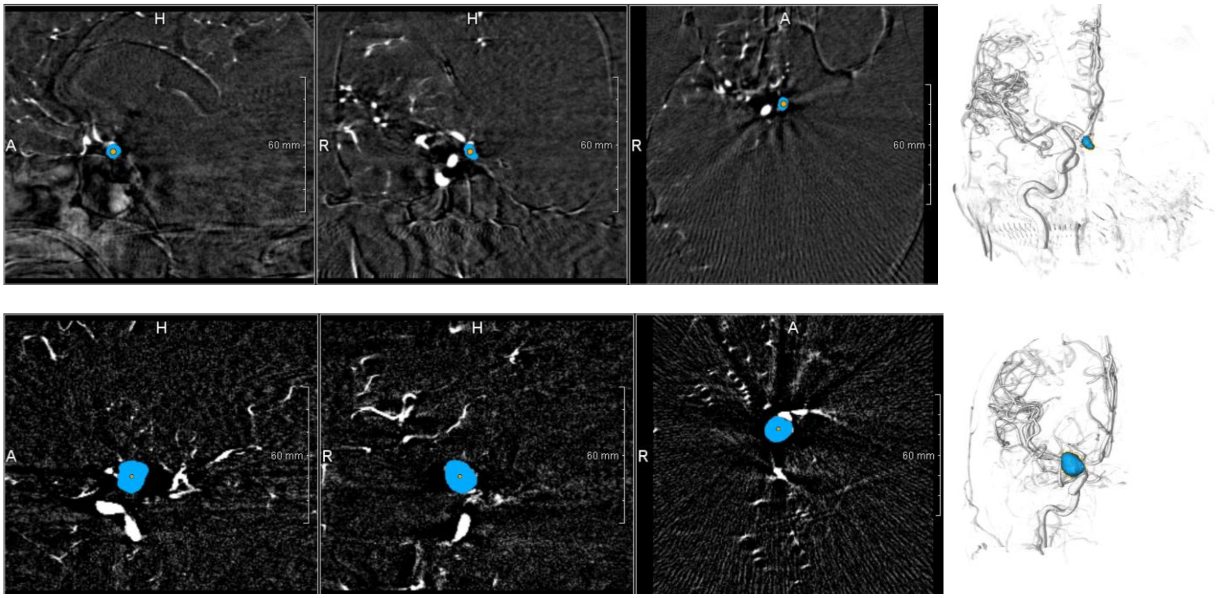


Figure 2: Example of datasets with an annotated aneurysm shown in orthogonal views in 2D as well as in a 3D volume rendering. The blue masks represent the segmentation, the orange point, the aneurysm center. Typical artifacts of rotation X-ray angiography, such as ray artifacts as well as differences in coverage of the datasets can be observed.

In addition to the image data, mask and stl-files are provided for the 142 segmented aneurysms. The aneurysm centers are provided via an XML-file. The rupture information is stored in a table listing the aneurysms together with their rupture state (58% not ruptured vs 42% ruptured).

The average volume per aneurysm is 0.391 ml, meaning that for image volumes between 280 and 2350 ml, there is a strong imbalance between foreground and background, which the participants have to consider when designing the preprocessing and training setup.

Data for Task 1

A training case consists of an image dataset showing a contrast-enhanced cerebrovascular vessel tree. Furthermore, segmentation masks (stl- and image files) are provided, so that the bounding boxes for the aneurysms can be calculated.

The test cases for generating results to be uploaded contain only the image data.

23 cases will be used as test cases. The 92 cases released to the participants can be freely separated by the participants (e.g. for cross validation)

Data for Task 2

A training case consists of an image dataset showing a contrast-enhanced cerebrovascular vessel tree. Furthermore, segmentation masks (stl- and image files) are provided, which are considered as ground truth.

The test cases for generating results to be uploaded automatically contain only image data. For the generation of semi-automatic and interactive segmentation results, the aneurysm bounding box positions for the test cases will be released later, to enable local initialization.

23 cases will be used as test cases. The 92 cases released to the participants can be freely separated by the participants (e.g. for cross validation)

Data for Task 3

A training case consists of an image dataset showing a contrast-enhanced cerebrovascular vessel tree. Furthermore, segmentation masks (stl- and image files) and the rupture state of each aneurysm are provided.

The test cases for generating the risk classification per aneurysm to be uploaded contain image data and segmentation masks.

The total number of cases is the number of annotated datasets available at the time of submission of this challenge proposal. We intend to increase this number for later events. The number of test cases amounts to 20% of the available datasets.

ASSESSMENT METHODS

Metrics

Task 1: Aneurysm Detection

For the aneurysm detection challenge, we expect participants to submit for each case per aneurysm a point coordinate representing the location $\vec{x}_{cA}$ of the aneurysm (ideally the center of the bounding box) in world coordinates. Additional two vectors $\vec{v}_{Ab1}$ and $\vec{v}_{Ab2}$ describe the orientation and extent of the bounding box. If for a list of points $X_c = \{\vec{x}_{c1}, \dots \vec{x}_{cn}\}$ for a case c at least one point coordinate corresponds to a voxel inside an aneurysm mask M_{cA} the aneurysm cA is considered detected: $\exists\, \vec{x}_{ci} \in Xc: \vec{x}_{ci} \subset M_{cA}$.

We intend to calculate the following metrics:

a) **Recall $\boldsymbol{R}$**(true positive rate, sensitivity):

$$R = \frac{\sum_{c\ in\ Cases} |\{cA|\ \exists\, \vec{x}_{ci} \in Xc: \vec{x}_{ci} \subset M_{cA}\}|}{\sum_{c\ in\ Cases} |\{cA\}|}$$

b) **Precision $\boldsymbol{P}$**(positive predictive value):

$$P = \frac{\sum_{c\ in\ Cases} |\{cA|\ \exists\, \vec{x}_{ci} \in Xc: \vec{x}_{ci} \subset M_{cA}\}|}{\sum_{c\ in\ Cases} |\{cA|\ \exists\, \vec{x}_{ci} \in Xc: \vec{x}_{ci} \subset M_{cA}\}| + \sum_{c\ in\ Cases} |\{\vec{x}_{ci} \in Xc|\ \nexists\, cA: \vec{x}_{ci} \subset M_{cA}\}|}$$

c) **Coverage $\boldsymbol{C_{cA}}$**of aneurysms cA by bounding boxes BB_{cA}

$$C_{cA} = \sum_{c\ in\ Cases}\ \sum_{cA\ in\ Aneurysms\ of\ c} |M_{cA} \setminus BB_{cA}|$$

d) **Bounding box fit $\boldsymbol{F_{cA}}$**(max distance of bounding box from mask along main axes of the bounding box)

$$F_{cA} = \sum_{c\ in\ Cases\ cA\ in\ Aneurysms\ of\ c} \sum \max_{\vec{x}_{ci} \in M_{cA}} \left(\min_{\vec{m}_j \in M_{cA}} \left(\left| \left(\vec{x}_{ci} + \vec{v}_{ib1} - \vec{m}_j \right) \right|, \left| \left(\vec{x}_{ci} + \vec{v}_{ib1} - \vec{m}_j \right) \right| \right) \right)$$

The major goal in detection is to make sure that aneurysms, which may pose a stroke risk, are not overlooked, so the sensitivity is an important measure. On the other hand, if the whole image is marked, the aneurysms would be included but the information would be meaningless, so the precision is important as well. A bounding box is helpful if it supports the visualization and postprocessing. To this end it should contain the aneurysm but be as small as possible.

Ranking will be based the $\boldsymbol{F_2}$**-score** that combines recall R and precision P considering recall twice as important as precision:

$$F_2 = 5\frac{PR}{4P + R}$$

The bounding box metrics will only be used in case of an equal ranking. Results will then be further ranked according to the aneurysm coverage C_{cA}, and the bounding box fit F_{cA} if this is not decisive.

The ranking score is chosen such that sensitivity is weighted stronger than precision, because missing a risk structure is considered worse than providing a false positive result. The coverage and bounding box fit are considered as second and third level measures for ranking of results with equal F2 score because they are considered less important.

The authors are expected to perform cross-validation on the training dataset themselves.

Task 2: Aneurysm Segmentation

The metrics for the segmentation assessment is based on the comparison segmentation of the masks M^*_{cA} provided by the participants with ground truth masks M_{cA} from the expert annotations. We intend to calculate standard metrics for segmentation results. Class probabilities will not be considered.

e) **Jaccard coefficient**:

$$J(M^*_{cA}, M_{cA}) = \frac{M^*_{cA} \cap M_{cA}}{M^*_{cA} \cup M_{cA}}$$

f) **Hausdorff distance**:

$$HD(M^*_{cA}, M_{cA}) = \max\left(\max_{\vec{m}^*_j \in M^*_{cA}} \left(\min_{\vec{m}_i \in M_{cA}} \left| \vec{m}^*_j - \vec{m}_i \right| \right) \right)$$

g) **Average distance**:

$$AVD(M^*_{cA}, M_{cA}) = \frac{1}{2|M^*_{cA}|} \sum_{\vec{m}^*_j \in M^*_{cA}} \min_{\vec{m}_i \in M_{cA}} \left| \vec{m}^*_j - \vec{m}_i \right| + \frac{1}{2|M_{cA}|} \sum_{\vec{m}_i \in M_{cA}} \min_{\vec{m}^*_j \in M^*_{cA}} \left| \vec{m}^*_j - \vec{m}_i \right|$$

h) **Pearson correlation coefficient** $\boldsymbol{r}$ between predicted $V^*$and reference volume V of all aneurysms

$$r = \frac{|\{cA\}| \sum_{\{cA\}} V^*_{cA} V_{cA} - \left(\sum_{\{cA\}} V^*_{cA}\right)\left(\sum_{\{cA\}} V_{cA}\right)}{\sqrt{|\{cA\}| \sum_{\{cA\}} {V^*_{cA}}^2 - \left(\sum_{\{cA\}} V^*_{cA}\right)^2} \sqrt{|\{cA\}| \sum_{\{cA\}} {V_{cA}}^2 - \left(\sum_{\{cA\}} V_{cA}\right)^2}}$$

i) **Bias (**$\boldsymbol{b}$**)** computed as the mean absolute difference of predicted and reference volume

$$b = \frac{1}{|\{cA\}|} \sum_{\{cA\}} |V_{cA}^{*} - V_{cA}|$$

j) **Standard deviation (σ)** of the difference between predicted and reference volumes.

The segmentation is the basis for the quantitative assessment of the aneurysms. It should enable the extraction of shape and volume parameters for the assessment of change over time or the comparison with decision thresholds. Therefore, the overlap, and distance from reference segmentations is important. For the assessment of volumes, we also analyze, how well the results correlate over the cohort and if there is a bias.

For the ranking, we will perform a normalization according to the maximum among all participants so that each individual metric takes a value between 0 (worst case among all participants) and 1 (perfect fit between the reference and predicted segmentation). The ranking score is calculated as the average of the normalized metrics. We consider all metrics as equally important for the application context and therefore try to integrate them with equal weight in the scoring system.

Task 3: Rupture risk

For the assessment of the rupture risk classification, we will calculate recall and precision. Ranking will be based on the $\boldsymbol{F_2}$**-Score**.
The rupture risk of an aneurysm should not be overlooked. On the other hand, too many false positives mean a tedious screening for the physician, who has to review the risk assessment for decision making. The F2-score combines recall and precision such that the identification of aneurysms at risk is considered more important than the avoidance of a false positive risk classification.

General Remarks

Algorithm runtime and, where applicable, interaction time are crucial parameters with regard to clinical applicability and must be provided together with **hardware requirements** for all submissions.

During the submission test phase, submissions will be checked for completeness, and participants will be notified if cases are missing. During the validation phase, missing cases will be interpreted as algorithm failure (worst possible metric value in each category). The metrics per submission will be shown separately in a table with accompanying boxplots so that it is possible to compare the algorithm performance per class separately and analyze biases. For each metric, we will analyze the coefficient of variation.
For Task 2, we will analyze which measure contributed most in the ranking process.

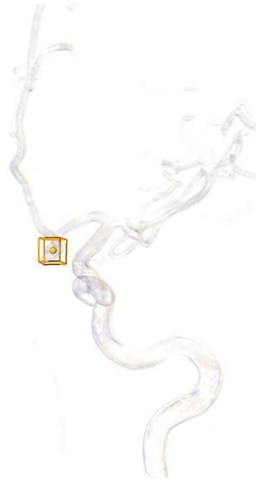


Figure 3: Example for centerpoint and bounding box for a detected aneurysm.