Intracranial Aneurysm Detection and Segmentation Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Intracranial Aneurysm Detection and Segmentation Challenge

Challenge acronym

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

ADAM: Aneurysm Detection And segMentation

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.

Introduction
Intracerebral aneurysms are found in 3% of the general population, and some groups have a higher risk. If an aneurysm ruptures it causes bleeding in the brain (subarachnoid haemorrhage). [1] Early detection of intracranial aneurysms, as well as accurate measurement and assessment of shape, is important in clinical routine. This enables careful monitoring of the growth and rupture risk of aneurysms to allow informed treatment decisions to be made [2]. Currently, contrast-enhanced computed tomography angiography scans (CTA) and non-contrast 3D time-of-flight magnetic resonance angiography (TOF-MRA) are the most common imaging techniques for this purpose.

However, intracranial aneurysm detection and measurement can sometimes be difficult – especially for small aneurysms [1]. It has been cited that about 10% of the aneurysms, mostly small ones, are still missed [3]. For small aneurysms (<5mm) it has been reported that detection by radiologists from MRAs can have a sensitivity as low as 35% [4].

Increased knowledge on risk factors for aneurysm presence, such a positive family history for the disease, has led to more individuals being preventively screened with MRA [5]. With more patients being screened, it is becoming important to reduce the clinical workflow duration, whilst still allowing the accurate detection and diagnosis of an aneurysm. Automatic methods of detection of aneurysms from TOF-MRAs would allow the speed of clinical workflow to be increased, without compromising accuracy.

Furthermore, automated volumetric segmentation would enable more reliable measurements and characteristics of aneurysms to be derived and considered for rupture risk prediction. For example, it is known that shape characteristics such as non-spherical and lobular shape are associated with elevated rupture risk [6, 7, 8]. Based on these shape features, and the associated rupture risk, a more informed treatment decision can be made. Shape of

an unruptured intracranial aneurysm can also have an effect on the treatment outcome of a patient. Shape features of the aneurysms, derived from volumetric segmentations, could further aid treatment complication prediction models [9].

Technical point of view
Various different (semi-) automatic methods for the detection and segmentation of intracranial aneurysms exist [10, 11]. Many detection methods are developed for CTA or Digital Subtraction Angiography (DSA) 2D images [12, 13]. However, in the clinic, MRI is best suited for regular follow-up as it requires neither intravenous contrast agent nor radiation. In addition, some treated (e.g. coiled) aneurysms can create large artefacts on CTA, so it is often necessary to assess for recanalization on MRA without artefacts. As TOF-MRA is increasingly used in clinical routine, characterisation and rupture risk assessment of aneurysms for MRA are becoming more important [14]. Hence, there is a need for accurate detection and segmentation methods from TOF-MRA. Aneurysms can be small, have very different shapes and occur at many different locations. In addition, fusiform widening of branching vessels can mimic small aneurysms. This leads to an exciting technical challenge to automatically detect and segment aneurysms, and includes generating creative and novel methods for medical image segmentation.

Impact
The purpose of this challenge is to automatically detect and segment intracranial aneurysms from TOF-MRA images. Automatic detection can aid a radiologist in diagnosis of intracranial aneurysms and will likely speed up the clinical workflow. Volumetric segmentation allows analysis of the size and shape of the aneurysms which may provide new biomarkers for use in rupture risk prediction models. Eventually, this may result in more informed decisions being made with regard to treatment of intracranial aneurysms.

Challenge keywords

List the primary keywords that characterize the challenge.

Detection, Segmentation, Aneurysms, MR angiography

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.

40 participating teams.
This estimate is made based on previous challenges by the same institute such as WMH segmentation challenge 2017 and MRBrainS18.

Publication and future plans

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

After the submission date and presentation of results at MICCAI 2020, the results will be summarised in a journal paper. All teams who submitted before the deadline and presented their results at MICCAI 2020 will be included in the paper. Each team is allowed two co-authors in this paper. Each participating team is welcome to publish their own results but we request that they cite the organisers' summary journal paper; or the challenge website in case the journal paper is not yet published. We also request that all teams notify the organisers of this challenge about any publication that is (partly) based on the results in order for us to maintain a list of publications associated with the challenge.
The results and our corresponding evaluation of all participating teams will be made publicly available on the created website.

Future Plans: The challenge will remain open after the MICCAI deadline and the Image Sciences Institute of the UMC Utrecht will continue to support it. This will provide a supported system for bench-marking of future proposed methods.

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 challenge will be run using the institutional website of the Image Sciences Institute (www.isi.uu.nl).
We have adequate computational resources for the evaluation of the challenge in our institute.
For the challenge session, we would require a standard room with a projector and screen.

TASK: Intracranial 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.

Introduction
Intracerebral aneurysms are found in 3% of the general population, and some groups have a higher risk. If an aneurysm ruptures it causes bleeding in the brain (subarachnoid haemorrhage). [1] Early detection of intracranial aneurysms, as well as accurate measurement and assessment of shape, is important in clinical routine. This enables careful monitoring of the growth and rupture risk of aneurysms to allow informed treatment decisions to be made [2]. Currently, contrast-enhanced computed tomography angiography scans (CTA) and non-contrast 3D time-of-flight magnetic resonance angiography (TOF-MRA) are the most common imaging techniques for this purpose.

However, intracranial aneurysm detection and measurement can sometimes be difficult – especially for small aneurysms [1]. It has been cited that about 10% of the aneurysms, mostly small ones, are still missed [3]. For small aneurysms (<5mm) it has been reported that detection by radiologists from MRAs can have a sensitivity as low as 35% [4].

Increased knowledge on risk factors for aneurysm presence, such a positive family history for the disease, has led to more individuals being preventively screened with MRA [5]. With more patients being screened, it is becoming important to reduce the clinical workflow duration, whilst still allowing the accurate detection and diagnosis of an aneurysm. Automatic methods of detection of aneurysms from TOF-MRAs would allow the speed of clinical workflow to be increased, without compromising accuracy.

Furthermore, automated volumetric segmentation would enable more reliable measurements and characteristics of aneurysms to be derived and considered for rupture risk prediction. For example, it is known that shape characteristics such as non-spherical and lobular shape are associated with elevated rupture risk [6, 7, 8]. Based on these shape features, and the associated rupture risk, a more informed treatment decision can be made. Shape of an unruptured intracranial aneurysm can also have an effect on the treatment outcome of a patient. Shape features of the aneurysms, derived from volumetric segmentations, could further aid treatment complication prediction models [9].

Technical point of view
Various different (semi-) automatic methods for the detection and segmentation of intracranial aneurysms exist [10, 11]. Many detection methods are developed for CTA or Digital Subtraction Angiography (DSA) 2D images [12, 13]. However, in the clinic, MRI is best suited for regular follow-up as it requires neither intravenous contrast agent nor radiation. In addition, some treated (e.g. coiled) aneurysms can create large artefacts on CTA, so it is often necessary to assess for recanalization on MRA without artefacts. As TOF-MRA is increasingly used in clinical routine, characterisation and rupture risk assessment of aneurysms for MRA are becoming more important [14]. Hence, there is a need for accurate detection and segmentation methods from TOF-MRA. Aneurysms can be small, have very different shapes and occur at many different locations. In addition, fusiform widening of branching vessels can mimic small aneurysms. This leads to an exciting technical challenge to automatically detect and

segment aneurysms, and includes generating creative and novel methods for medical image segmentation.

Impact
The purpose of this challenge is to automatically detect and segment intracranial aneurysms from TOF-MRA images. Automatic detection can aid a radiologist in diagnosis of intracranial aneurysms and will likely speed up the clinical workflow. Volumetric segmentation allows analysis of the size and shape of the aneurysms which may provide new biomarkers for use in rupture risk prediction models. Eventually, this may result in more informed decisions being made with regard to treatment of intracranial aneurysms.

Keywords

List the primary keywords that characterize the task.

detection, aneurysms, MR angiography

ORGANIZATION

Organizers

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

Kimberley Timmins, Image Sciences Institute, UMC Utrecht, the Netherlands (Main organiser)
Edwin Bennink, Image Sciences Institute, UMC Utrecht, the Netherlands
Irene van der Schaaf, Department of Radiology, UMC Utrecht, the Netherlands
Birgitta Velthuis, Department of Radiology, UMC Utrecht, the Netherlands
Ynte Ruigrok, Department of Neurology, UMC Utrecht, the Netherlands
Hugo Kuijf, Image Sciences Institute, UMC Utrecht, the Netherlands

b) Provide information on the primary contact person.

Kimberley Timmins, Image Sciences Institute, UMC Utrecht, the Netherlands
k.m.timmins@umcutrecht.nl

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

One-time event with a fixed submission deadline. Submissions received before the deadline will be considered for the awards at the MICCAI challenge session.

The challenge will remain open after the MICCAI deadline and the Image Sciences Institute of the UMC Utrecht will continue to support it. This will provide a supported system for bench-marking of future proposed methods.

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.

Institutional website of the Image Sciences Institute (www.isi.uu.nl).

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

The URL will be adam.isi.uu.nl

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.

The data for training any algorithms is not restricted to the data provided by the challenge. Public datasets may be used but their usage should be mentioned in the methods sections. We will also allow the use of private datasets as long as the authors are willing to share this dataset with others for evaluation purposes.

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

May not participate.

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

A certificate and small gift will be provided to the top teams that have submitted before the deadline and are ranked in the top 3 for each task.

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.

Result Announcement
The results will be announced publicly at the MICCAI 2020 challenge session. After the session, the results will be published on the challenge website.

Conference Session
The top 3 and 4 to 10 ranking teams will be advised two weeks prior to the MICCAI challenge meeting but they will not receive their individual ranking. The top 3 teams will be asked to prepare a 10-minute presentation of their method at the challenge session. The remaining 7 top ranked teams (out of the 10) will be asked to prepare a short pitch (1 minute) about their method for the meeting. This will be performed separately for the two tasks: detection and segmentation. If any of the top teams overlap in both tasks, then they will not be required to present twice. All participating teams in the challenge that submit their methods before the deadline will be asked to prepare a poster for the challenge session. After all presentations at the challenge meeting, all rankings will be announced including the individual rankings of the top-10 and the winner of the challenge for each task.

Post-Challenge
After the challenge, the website will be updated with the results. Since the challenge will be kept open, a live leader board will be added on the website for submissions received after the deadline. This can be used for benchmarking of future submitted methods.

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).

After the submission date and presentation of results at MICCAI 2020, the results will be summarised in a journal paper. All teams who submitted before the deadline and presented their results at MICCAI 2020 will be included in the paper. Each team is allowed two co-authors in this paper. Each participating team is welcome to publish their own results but we request that they cite the organisers' summary journal paper; or the challenge website in case the journal paper is not yet published. There is no embargo time defined. We also request that all teams notify the organisers of this challenge about any publication that is (partly) based on the results in order for us to maintain a list of publications associated with the challenge.

The results and our corresponding evaluation of all participating teams will be made publicly available on the challenge website after the conference session.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants will containerise their algorithms with Docker and submit these to the organisers. Detailed instructions and easy-to-follow examples will be provided on the website (example: https://wmh.isi.uu.nl/methods/) and, if needed, the organisers will help with containerisation. The organisers will run the submitted methods on the test data within their own institute using publicly available evaluation code. This guarantees that the test data remains secret and cannot be included in the training procedure of the techniques. If technical issues or bugs occur in the evaluation, we allow teams to submit a fix. This workflow has proven successful for previous MICCAI challenges at the Image Sciences Institute, such as MRBrainS18 and WMH Segmentation challenge 2017.

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.

Teams can evaluate their methods on the training data (by splitting it into separate training and validation sets) or on private data of their own institute. The evaluation code that will be used for the official evaluation on the test set will be made available to all teams via GitHub and the challenge website.
Teams are not able to evaluate their own results on the official test set. The test data will be kept secret and not released to the challenge participants. It is not possible to submit multiple times as the test set and the results are kept secret until the challenge session. Evaluation will be performed at the organiser's institute on the secret test data.
In case of technical issues with the method (e.g. related to containerisation), we allow participants to submit fixes.

After the MICCAI 2020 challenge session, the challenge will remain open for new submissions and updates of previously submitted methods.

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

Preliminary Timetable
24th February 2020: Date of planned proposal acceptance notification
3rd April 2020: Website online and Data Release
15th August 2020: Deadline for submission of methods
20th September 2020: Top 10 performing teams for each task contacted and asked to prepare oral presentation for conference. (top 3 (for each task) – 10minutes, remaining 7 (for each task) – 1 minute pitches)
4th-8th October 2020: Challenge session at MICCAI 2020 and release of all results.

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).

Not required and all data used will be anonymised.

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)

We will use the same Terms of Participation as previous challenges that we (and others) organised, e.g.: https://wmh.isi.uu.nl/wp-content/uploads/2019/04/agreement.pdf . Participants can register for the challenge and download the data to be used for a challenge submission. We ask that participants do not distribute the dataset, but instead refer others to the challenge website; so we can keep track of the data usage statistics. Data can be used for other purposes (i.e. other than preparing a challenge submission) after contacting the challenge organisers.

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 evaluation software, including code of how the rankings will be assessed, will be available on the challenge website as well as on GitHub. This will be similar as performed for the WMH challenge and MRBrainS. https://github.com/hjkuijf/wmhchallenge/blob/master/evaluation.py

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

The organisers will ask for approval from the participating teams to publish the team's containers on Docker Hub after the deadline. This is similar the WMH challenge and MRBrainS: https://hub.docker.com/u/wmhchallenge

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 have requested no direct funding for the challenge, are not sponsored and everything is paid for by our organisation. We have received a grant from Applied Data Science, Utrecht University to fund a second rater for annotations of the scans.

The access to the test data is limited to UMC Utrecht employees only, and all UMC Utrecht employees are excluded from participation in the challenge. The participating teams will not have access to the test data and all evaluation on test data will be performed by the organising team in house. Hence, we do not foresee any conflicts of interest.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, Intervention planning, Screening, Diagnosis, 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 is any adult patient undergoing a clinical brain TOF-MRA with potential intracranial aneurysms. Some MRA scans will be negative (i.e. a patient without any diagnosed intracranial aneurysms) to reflect the clinical setting. They may be scanned for the following reasons: (1) symptoms associated with aneurysm development; (2) follow-up scans of patients already known to have an intracranial aneurysm or patients who have already undergone intracranial aneurysm treatment (may have further aneurysms); (3) Patients screened for positive family history. Two examples of positive family history studies underway at the UMC Utrecht are: (1) FAMSAB in which persons are screened for aneurysms because of a positive family history for ruptured aneurysms; and (2) ERASE where persons are screened for aneurysms because of a positive family history for unruptured aneurysms.

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

The challenge cohort consists of subjects with aneurysms who are a subset of patients taken from the cohort used in previous aneurysm growth study ELAPSS, [6] at UMC Utrecht, who had an available TOF-MRA. They were acquired retrospectively from clinical data. Subjects who had a diagnosed aneurysm were also acquired from ongoing studies to assess positive family history of aneurysms, at the UMC Utrecht (FAMSAB and ERASE as detailed in 16 a)).
Subjects without any diagnosed aneurysms were a subset of patients acquired from the ERASE or FAMSAB study database.
A total of 250 scans will be included, which includes 269 untreated aneurysms. All scans were taken from 2001-2019. The subjects with intracranial aneurysms had a median age of 54 (range 24-56), with 75% of subjects being female. A subset of the dataset includes two scans from the same subject, both a baseline and a >6 month follow-up scan, to reflect the real clinical data. The aneurysms ranged in size, with a range of maximum diameter from (0.7mm –15.9mm). 18% of the scans contain multiple aneurysms and 7.5% of the scans contained treated

aneurysms.

Imaging modality(ies)

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

Time of Flight Magnetic Resonance Angiography (TOF-MRA) and T2/T1 weighted MR imaging using 1T, 1.5T and 3T MRI scanners.

Context information

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

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

For every subject we will provide additional information along with the images. These are:
1) a text file with the 3D coordinates of the centre of mass of the aneurysms and the maximum radius of the aneurysm
2) a label image with labels:
0 = background
1 = untreated, unruptured aneurysm
2 = treated aneurysm
(note: label 2 will be ignore during evaluation)
3) Results of pre-processing, including:
a) The original TOF-MRA image and structural image
b) Co-registered structural image (registered to TOF-MRA)
c) Transformation parameters used for co-registration
d) Mask used to remove the face for anonymisation (if required)
All information will be provided in the image coordinates of the TOF-MRA.

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

The scans will be anonymised so no additional information regarding the patient, other than that described above, will be 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.

In both the target and challenge cohort, the image data will be acquired of the brain using TOF-MRA and T1/T2 weighted MRI.

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

Untreated, unruptured Intracranial Aneurysms

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)).

Sensitivity.

Additional points: Detection: To detect untreated, unruptured intracranial aneurysms from TOF-MRA images to aid a radiologist when examining a MRA for any aneurysms. Here, it is important to have a detection method with a high sensitivity to ensure all possible aneurysms are detected and none are missed. The second priority would be low false positive count, the more false positives that there are the longer it takes for a radiologist to screen the predicted aneurysm candidates. It is, therefore, a balance of both sensitivity and false positive count that will aid and increase the speed of the detection procedure by a radiologist.

The algorithm output should be a 3D coordinate (x,y,z) of the voxel at the binary centre of mass of the aneurysm. The coordinates should be in the same image coordinates as the original TOF MRA. Positive detection will be determined based on the predicted coordinates (x,y,z) being located within the maximum aneurysm radius based on the manual aneurysm mask.

Detection of the treated (e.g. coiled) aneurysms will not be considered when assessing the performance of untreated, unruptured aneurysm detection or segmentation. That is: any false positive detections at the location of treated aneurysms will be ignored during evaluation.

For both tasks, the description of desired output will be provided to all participants with a full in-depth explanation and examples provided on the challenge website. Furthermore, the aneurysm mask, centre of mass and radius are provided for each training case so the participant can self-evaluate based on this. The source code that will be used for the final evaluation will also be provided to participants.

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).

A variety of Philips MRI scanners (1.0, 1.5 and 3T) were used to acquire the TOF-MRA and T1/T2 images. An in-house developed software was used to develop the labels of the aneurysms from the TOF-MRA images.

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

Due to the clinical nature of this data set and that it was taken from multiple studies across many years, there is no set protocol or acquisition used across cases. This provides a diverse and realistic data set in which standard clinical protocols and routine were used.

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.

University Medical Center (UMC) Utrecht, The Netherlands

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).

Not applicable

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases both represent a Time of Flight MRA of a human brain and a corresponding structural image (either a T1 or T2 weighted MRI). Training and test cases will both have corresponding annotations of the aneurysm based on the TOF-MRA to use for training or evaluation of the method.
The desired result is a detection file with locations of the aneurysm in the TOF-MRA.

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

Each case consists of one TOF-MRA and one structural image.

Train Dataset: 110 cases
1. 90 cases containing at least one untreated, unruptured intracranial aneurysm
-> 35 baseline and 35 follow-up of the same subject
-> 20 unique subjects
2. 20 scans of subjects without intracranial aneurysms

Test Dataset (not released): 140 cases
1. 115 cases containing at least one untreated, unruptured intracranial aneurysm
-> 45 baseline and 45 follow-up of the same subject
-> 25 unique subjects
2. 25 cases of subjects without intracranial aneurysms

We are not providing a specific validation set and it is up to the participants to decide a validation/train set split based on the provided training data. Follow-up scans were performed at least 6 months after the baseline scan.

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

The number of training data was chosen as this allowed enough variability between scans and subjects to give a

broad, diverse dataset representing a dataset in a real clinical setting. Previous studies based on similar detection/segmentation methods suggest that approximately 100 scans of subjects containing aneurysms would be an adequate amount to allow detection/segmentation to be performed [7,13,14]. The size of the test data set allows the methods to be tested on a data set similar to that of a real-life dataset from the clinic. The test set also allows us, as the organising team, to perform detailed analysis on the different submitted methods and the variability of results.

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 scans used were real clinical data taken from a long time period 2001-2019. The structural image provided for each scan is either a T1 or T2 weighted MRI, depending on the protocol used. This means it includes a variety of different scanners, protocol parameters, resolutions and qualities. This represents a realistic clinical dataset, which makes the challenge more applicable and applicable to everyday clinical use.

The subjects with two scans of both baseline and follow-up represents real world data that would come from the clinic. It is normal practise that a patient with an unruptured aneurysm, considered not at risk of rupture, to undergo follow-up imaging some time later. It also allows the organisers to perform evaluation on how the algorithm works within scans of the same subject.

The training data will include scans with a distribution of different scan protocol parameters and different aneurysm characteristics. We will ensure that the training data is representative of the test data, by controlling the split according to important parameters such as aneurysm location and size, scanning protocol parameters and the age/sex of the patient. No subjects are in both test and training set. This will ensure that all types of aneurysms and images are present in both the training and test data sets, and that the training set is representative of the test set.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Manual segmentations were made on axial slices of the TOF-MRAs using in-house developed software. These were then converted to give a 3D binary mask of the aneurysm. The annotations were performed by two raters. Both raters were given training on the annotation software by the software developers.

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 protocol for aneurysm annotation was developed by an interventional neuro-radiologist. On each axial slice of the TOF-MRA a contour around the outline of the aneurysm was drawn. The annotations were always drawn to be from the level of the neck to the dome of the aneurysm. The neck corresponds to the opening of the aneurysm from the parent vessel. The dome is the furthest part of the aneurysm from the parent vessel. None of the parent vessel was included in the annotation. During annotation, the raters had access to the structural image and a radiologist report made at the time of the scan. The radiologists' reports indicated the rough location and size of

the aneurysm.

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.

Aneurysms were diagnosed in all images as part of clinical routine. Upon inclusion of subjects in the ELAPSS/FAMSAB/ERASE study, an experienced interventional neuro-radiologist with more than 10 years of experience in the field manually detected each aneurysm and provided a detailed description and location. This neuro-radiologist trained a second rater with extensive experience in medical image analysis and the annotation software, but not specifically intracranial aneurysms. Once this second rater was on par with the first rater (as assessed by comparing intra/inter rater variability), the rater annotated all images in the dataset. Finally, the first and second rater will assess the full dataset together and make any required modifications in consensus. Hence, this will result in a consensus annotation that will be the official ground truth data set. The subset of annotations that have been annotated independently by both raters will be used to determine interrater variability.

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 multiple annotations will only be used to determine interrater variability. The actual annotations used will be those decided in consensus as described above.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The data will be pre-processed using n4 bias field correction [17], the face will be masked, and the structural image will be co-registered with the TOF-MRA. Both the original anonymised and the pre-processed data will be provided, alongside the face-mask used. This allows participants to implement their own pre-processing if they desire. A structural image (either T1 or T2) has been provided for each TOF-MRA and participants are welcome to use these as desired whether for pre-processing for inclusion in the actual training data. For the structural images, both the original image and a co-registered image will be provided, including the transformation parameters. The test data will not be provided to participants, but will undergo the exact same pre-processing as for the training data.

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.

Current findings, based on manual annotations made on a subset of the data-set by an experienced rater and a student, suggest that the manual annotations may vary by approximately 5% in volume. We have obtained a grant from Focus area Applied Data Science, Utrecht University to fund a second rater who will be trained by the experienced neuro-radiologist. Thus determination of the true magnitude of these errors will be in the spring of 2020 and be released on the challenge website. Based on a previous study using a similar manual segmentation technique [18] good inter- and intra-observer reliability of the volume was found with an intra-class correlation coefficient of 0.94.

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

The quality of the scans influences the variability of annotations between scans. However, the diverse range of scans with different qualities and resolutions gives a real representation of a true clinical dataset. The size of the aneurysm also influences the annotation variability, as smaller aneurysms are understandably harder to annotate and voxel size has more of an influence.

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)

Detection:
- Sensitivity
- False Positive Count (total per scan)

Indication of how this metrics can be determined can be found here: [19]

These metrics will be determined for each scan in the test set. For the final evaluation, a mean of each of the metrics over all of the test scans will be determined. For all metrics, other than false positive count, the mean of the metrics will be weighted to the number of aneurysms per scan. This allows for evaluation of the performance of the method across all of the scans and aneurysms which is important due to the variance between scans.

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

Detection: The chosen metrics are similar to those used in previous similar detection challenges [20], as well as those used in clinical practice.[4]

Sensitivity – Sensitivity measures the true positive rate, i.e. the number of detected aneurysms that correspond to true aneurysms as a ratio of all detected aneurysms. By maximising the sensitivity, it means that more aneurysms are being detected overall. As a clinician it is preferable that more aneurysm candidates are identified, and then the incorrect detections can be manually removed. If the sensitivity is low, then it is possible that more aneurysms are being missed. The sensitivity is an indication of how good the detection system is.

False Positive Count (total per scan) – The total number of false positives in each scan allows for identification of how many detections made by the method do not correspond to true aneurysms. Although it is important that all aneurysms are detected, it is also important that there are not too many false positives detected such that the detection tool does not actually aid the radiologist. This metric is used to balance the sensitivity of the detection tool. The false positive count will be determined for each scan. The false positive count will be log transformed before determining the ranking, in order to accurately assess small/large differences between values.

We note that specifically AUC and precision will not be used to assess the detection task as we do not want to assess the relative amount of false positives to true positives. Instead, we are interested that all aneurysms are detected, so the false positive detection count alone is important. True positives are already included in the sensitivity metric, and we do not need to assess this twice.

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.

Each metric, defined in item 26 a), is averaged over all test scans. For each metric, the participating teams are sorted from best to worst. The best team receives a rank of 0 and the worst team a rank of 1; all other teams are ranked (0 -1) relative to their performance within the range of that metric. Finally, the ranks for each metric are averaged into the overall rank that is used for the results.

For example: the best team A has a sensitivity of 80% and the worst team B a sensitivity of 60%. In the ranking: A=0.00 and B=1.00. Another team C has a sensitivity of 78%, which is then ranked at 1.0 - (78 - 60) / (80 - 60) = 0.10.

The metric ranges (e.g. 80 - 60) will be fixed for all submissions after MICCAI 2020. In that way future scores will be relative to the MICCAI 2020 scores.

This is the same ranking system used the WMH challenge, which worked well. It allows a fair and direct comparison of the evaluation of the different methods. The variance between scans is considered, by averaging across all test scans.

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

Not Applicable. No submissions will have missing results as we are performing the evaluation on test sets in house, using the code provided with Docker containers. If algorithm does not produce any output it will be considered as “no aneurysms detected”. If we notice that a method has technical errors that prevent it from running we will contact the participants to have this fixed.

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

The described ranking scheme will be used to give equal weighting to all metrics (when they may have different ranges) and allow a fair comparison between teams. This has worked well for the WMH and MRBrainS challenges. The relative ranking clearly shows the small/large performance differences between competing methods; which does not show very well in conventional ranking (e.g. methods ranked 1 and 10 may be closer in performance than methods 13 and 16).

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.

No data will be missing as the test data will not be provided to participants and all evaluation is done at the organisers’ institute. Evaluation will be performed using the metrics and ranking as described in the previous section (26). The evaluation and ranking code will be published online beforehand.
The teams will be ranked from best to worst on an average of the ranking of all the metrics. 95% confidence

intervals on each individual metric and on the final ranking will be computed using bootstrapping. This will be performed by randomly taking a large number of samples from the test set using sampling by replacement. Non-overlapping confidence intervals, determined in this way, indicate a significant difference between the methods.

Full statistical analysis of all methods will be detailed in the final journal paper after the conference.

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

95% confidence intervals and bootstrapping were used to ensure that the methods were statistically different from each other with regard to each metric.

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.

In further analyses, algorithms will be combined using an ensemble method in the journal paper. For example, for the detection task combining of methods may be performed by considering the proximity of the predicted coordinates of each method relative to each other. Methods can be combined based on these results and the resulting co-ordinates evaluated and ranked separately. A similar ensemble method has been used in previous detection challenges. [20]

A ranking will also be produced based on the intra-subject performance to determine those methods that perform well to assess growth. This will be defined by investigating if the methods are capable of identifying growth, where the growth is a change of volume of the segmented aneurysm which exceeds the volumetric similarity. This will be compared to growth as it is clinically defined and differences investigated.

As both a TOF-MRA and a structural image (T1 or T2) are provided, analysis will also be performed on the modality of image used for each of the methods and how this affected the results and ranking. Analysis of the submitted algorithm performance compared to the interrater reproducibility will also be measured.

Analyses will also be performed on the different types and structures of methods used, however this analysis will be dependent on the type of methods submitted (e.g. machine-learning vs classical, different deep-learning techniques etc.) This may be performed in a similar way as in the WMH challenge paper [21].

Full statistical analysis of all methods will be detailed in the final journal paper after the conference.

TASK: Intracranial 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.

Introduction
Intracerebral aneurysms are found in 3% of the general population, and some groups have a higher risk. If an aneurysm ruptures it causes bleeding in the brain (subarachnoid haemorrhage). [1] Early detection of intracranial aneurysms, as well as accurate measurement and assessment of shape, is important in clinical routine. This enables careful monitoring of the growth and rupture risk of aneurysms to allow informed treatment decisions to be made [2]. Currently, contrast-enhanced computed tomography angiography scans (CTA) and non-contrast 3D time-of-flight magnetic resonance angiography (TOF-MRA) are the most common imaging techniques for this purpose.

However, intracranial aneurysm detection and measurement can sometimes be difficult – especially for small aneurysms [1]. It has been cited that about 10% of the aneurysms, mostly small ones, are still missed [3]. For small aneurysms (<5mm) it has been reported that detection by radiologists from MRAs can have a sensitivity as low as 35% [4].

Increased knowledge on risk factors for aneurysm presence, such a positive family history for the disease, has led to more individuals being preventively screened with MRA [5]. With more patients being screened, it is becoming important to reduce the clinical workflow duration, whilst still allowing the accurate detection and diagnosis of an aneurysm. Automatic methods of detection of aneurysms from TOF-MRAs would allow the speed of clinical workflow to be increased, without compromising accuracy.

Furthermore, automated volumetric segmentation would enable more reliable measurements and characteristics of aneurysms to be derived and considered for rupture risk prediction. For example, it is known that shape characteristics such as non-spherical and lobular shape are associated with elevated rupture risk [6, 7, 8]. Based on these shape features, and the associated rupture risk, a more informed treatment decision can be made. Shape of an unruptured intracranial aneurysm can also have an effect on the treatment outcome of a patient. Shape features of the aneurysms, derived from volumetric segmentations, could further aid treatment complication prediction models [9].

Technical point of view
Various different (semi-) automatic methods for the detection and segmentation of intracranial aneurysms exist [10, 11]. Many detection methods are developed for CTA or Digital Subtraction Angiography (DSA) 2D images [12, 13]. However, in the clinic, MRI is best suited for regular follow-up as it requires neither intravenous contrast agent nor radiation. In addition, some treated (e.g. coiled) aneurysms can create large artefacts on CTA, so it is often necessary to assess for recanalization on MRA without artefacts. As TOF-MRA is increasingly used in clinical routine, characterisation and rupture risk assessment of aneurysms for MRA are becoming more important [14]. Hence, there is a need for accurate detection and segmentation methods from TOF-MRA. Aneurysms can be small, have very different shapes and occur at many different locations. In addition, fusiform widening of branching vessels can mimic small aneurysms. This leads to an exciting technical challenge to automatically detect and

segment aneurysms, and includes generating creative and novel methods for medical image segmentation.

Impact
The purpose of this challenge is to automatically detect and segment intracranial aneurysms from TOF-MRA images. Automatic detection can aid a radiologist in diagnosis of intracranial aneurysms and will likely speed up the clinical workflow. Volumetric segmentation allows analysis of the size and shape of the aneurysms which may provide new biomarkers for use in rupture risk prediction models. Eventually, this may result in more informed decisions being made with regard to treatment of intracranial aneurysms.

Keywords

List the primary keywords that characterize the task.

detection, segmentation, aneurysms, MR angiography

ORGANIZATION

Organizers

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

Kimberley Timmins, Image Sciences Institute, UMC Utrecht, the Netherlands (Main Organiser)
Edwin Bennink, Image Sciences Institute, UMC Utrecht, the Netherlands
Irene van der Schaaf, Department of Radiology, UMC Utrecht, the Netherlands
Birgitta Velthuis, Department of Radiology, UMC Utrecht, the Netherlands
Ynte Ruigrok, Department of Neurology, UMC Utrecht, the Netherlands
Hugo Kuijf, Image Sciences Institute, UMC Utrecht, the Netherlands

b) Provide information on the primary contact person.

Kimberley Timmins, Image Sciences Institute, UMC Utrecht, the Netherlands
k.m.timmins@umcutrecht.nl

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

One-time event with a fixed submission deadline. Submissions received before the deadline will be considered for the awards at the MICCAI challenge session.

The challenge will remain open after the MICCAI deadline and the Image Sciences Institute of the UMC Utrecht will continue to support it. This will provide a supported system for bench-marking of future proposed methods.

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.

Institutional website of the Image Sciences Institute (www.isi.uu.nl).

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

The URL will be adam.isi.uu.nl

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.

The data for training any algorithms is not restricted to the data provided by the challenge. Public datasets may be used but their usage should be mentioned in the methods sections. We will also allow the use of private datasets as long as the authors are willing to share this dataset with others for evaluation purposes.

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

May not participate.

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

A certificate and small gift will be provided to the top teams that have submitted before the deadline and are ranked in the top 3 for each task.

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.

Result Announcement
The results will be announced publicly at the MICCAI 2020 challenge session. After the session, the results will be published on the challenge website.

Conference Session
The top 3 and 4 to 10 ranking teams will be advised two weeks prior to the MICCAI challenge meeting but they will not receive their individual ranking. The top 3 teams will be asked to prepare a 10-minute presentation of their method at the challenge session. The remaining 7 top ranked teams (out of the 10) will be asked to prepare a short pitch (1 minute) about their method for the meeting. This will be performed separately for the two tasks: detection and segmentation. If any of the top teams overlap in both tasks, then they will not be required to present twice. All participating teams in the challenge that submit their methods before the deadline will be asked to prepare a poster for the challenge session. After all presentations at the challenge meeting, all rankings will be announced including the individual rankings of the top-10 and the winner of the challenge for each task.

Post-Challenge
After the challenge, the website will be updated with the results. Since the challenge will be kept open, a live leader board will be added on the website for submissions received after the deadline. This can be used for benchmarking of future submitted methods.

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).

After the submission date and presentation of results at MICCAI 2020, the results will be summarised in a journal paper. All teams who submitted before the deadline and presented their results at MICCAI 2020 will be included in the paper. Each team is allowed two co-authors in this paper. Each participating team is welcome to publish their own results but we request that they cite the organisers' summary journal paper; or the challenge website in case the journal paper is not yet published. There is no embargo time defined. We also request that all teams notify the organisers of this challenge about any publication that is (partly) based on the results in order for us to maintain a list of publications associated with the challenge.

The results and our corresponding evaluation of all participating teams will be made publicly available on the challenge website after the conference session.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants will containerise their algorithms with Docker and submit these to the organisers. Detailed instructions and easy-to-follow examples will be provided on the website (example: https://wmh.isi.uu.nl/methods/) and, if needed, the organisers will help with containerisation. The organisers will run the submitted methods on the test data within their own institute using publicly available evaluation code. This guarantees that the test data remains secret and cannot be included in the training procedure of the techniques. If technical issues or bugs occur in the evaluation, we allow teams to submit a fix. This workflow has proven successful for previous MICCAI challenges at the Image Sciences Institute, such as MRBrainS18 and WMH Segmentation challenge 2017.

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.

Teams can evaluate their methods on the training data (by splitting it into separate training and validation sets) or on private data of their own institute. The evaluation code that will be used for the official evaluation on the test set will be made available to all teams via GitHub and the challenge website.
Teams are not able to evaluate their own results on the official test set. The test data will be kept secret and not released to the challenge participants. It is not possible to submit multiple times as the test set and the results are kept secret until the challenge session. Evaluation will be performed at the organiser's institute on the secret test data.
In case of technical issues with the method (e.g. related to containerisation), we allow participants to submit fixes.

After the MICCAI 2020 challenge session, the challenge will remain open for new submissions and updates of previously submitted methods.

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

Preliminary Timetable
24th February 2020: Date of planned proposal acceptance notification
3rd April 2020: Website online and Data Release
15th August 2020: Deadline for submission of methods
20th September 2020: Top 10 performing teams for each task contacted and asked to prepare oral presentation for conference. (top 3 (for each task) – 10minutes, remaining 7 (for each task) – 1 minute pitches)
4th-8th October 2020: Challenge session at MICCAI 2020 and release of all results.

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).

Not required and all data used will be anonymised.

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)

We will use the same Terms of Participation as previous challenges that we (and others) organised, e.g.: https://wmh.isi.uu.nl/wp-content/uploads/2019/04/agreement.pdf . Participants can register for the challenge and download the data to be used for a challenge submission. We ask that participants do not distribute the dataset, but instead refer others to the challenge website; so we can keep track of the data usage statistics. Data can be used for other purposes (i.e. other than preparing a challenge submission) after contacting the challenge organisers.

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 evaluation software, including code of how the rankings will be assessed, will be available on the challenge website as well as on GitHub. This will be similar as performed for the WMH challenge and MRBrainS. https://github.com/hjkuijf/wmhchallenge/blob/master/evaluation.py

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

The organisers will ask for approval from the participating teams to publish the team's containers on Docker Hub after the deadline. This is similar the WMH challenge and MRBrainS: https://hub.docker.com/u/wmhchallenge

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 have requested no direct funding for the challenge, are not sponsored and everything is paid for by our organisation. We have received a grant from Applied Data Science, Utrecht University to fund a second rater for annotations of the scans.

The access to the test data is limited to UMC Utrecht employees only, and all UMC Utrecht employees are excluded from participation in the challenge. The participating teams will not have access to the test data and all evaluation on test data will be performed by the organising team in house. Hence, we do not foresee any conflicts of interest.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, Intervention planning, Screening, Diagnosis, CAD.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

The target cohort is any adult patient undergoing a clinical brain TOF-MRA with potential intracranial aneurysms. Some MRA scans will be negative (i.e. a patient without any diagnosed intracranial aneurysms) to reflect the clinical setting. They may be scanned for the following reasons: (1) symptoms associated with aneurysm development; (2) follow-up scans of patients already known to have an intracranial aneurysm or patients who have already undergone intracranial aneurysm treatment (may have further aneurysms); (3) Patients screened for positive family history. Two examples of positive family history studies underway at the UMC Utrecht are: (1) FAMSAB in which persons are screened for aneurysms because of a positive family history for ruptured aneurysms; and (2) ERASE where persons are screened for aneurysms because of a positive family history for unruptured aneurysms.

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

The challenge cohort consists of subjects with aneurysms who are a subset of patients taken from the cohort used in previous aneurysm growth study ELAPSS, [6] at UMC Utrecht, who had an available TOF-MRA. They were acquired retrospectively from clinical data. Subjects who had a diagnosed aneurysm were also acquired from ongoing studies to assess positive family history of aneurysms, at the UMC Utrecht (FAMSAB and ERASE as detailed in 16 a)).
Subjects without any diagnosed aneurysms were a subset of patients acquired from the ERASE or FAMSAB study database.
A total of 250 scans will be included, which includes 269 untreated aneurysms. All scans were taken from 2001-2019. The subjects with intracranial aneurysms had a median age of 54 (range 24-56), with 75% of subjects being female. A subset of the dataset includes two scans from the same subject, both a baseline and a >6 month follow-up scan, to reflect the real clinical data. The aneurysms ranged in size, with a range of maximum diameter from (0.7mm –15.9mm). 18% of the scans contain multiple aneurysms and 7.5% of the scans contained treated

aneurysms.

Imaging modality(ies)

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

Time of Flight Magnetic Resonance Angiography (TOF-MRA) and T2/T1 weighted MR imaging using 1T, 1.5T and 3T MRI scanners.

Context information

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

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

For every subject we will provide additional information along with the images. These are:
1) a text file with the 3D coordinates of the centre of mass of the aneurysms and the maximum radius of the aneurysm
2) a label image with labels:
0 = background
1 = untreated, unruptured aneurysm
2 = treated aneurysm
(note: label 2 will be ignore during evaluation)
3) Results of pre-processing, including:
a) The original TOF-MRA image and structural image
b) Co-registered structural image (registered to TOF-MRA)
c) Transformation parameters used for co-registration
d) Mask used to remove the face for anonymisation (if required)
All information will be provided in the image coordinates of the TOF-MRA.

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

The scans will be anonymised so no additional information regarding the patient, other than that described above, will be 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.

In both the target and challenge cohort, the image data will be acquired of the brain using TOF-MRA and T1/T2 weighted MRI.

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

Untreated, unruptured Intracranial Aneurysms

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)).

Segmentation: To segment untreated, unruptured intracranial aneurysms by providing a binary mask in the image coordinates of the original TOF-MRA. An ideal mask should have a Dice score and volumetric similarity that reflects the inter- and intra- variability of experts' manual annotations. This ensures the accurate volumetric assessment of aneurysms which is important for assessing growth and rupture risk. The boundary of the segmentation is also important for shape characterisation of the aneurysm and hence this will be assessed using the modified Hausdorff distance.

The algorithm output should be a binary mask of the predicted segmented aneurysm in the same image space as the original TOF MRA.

Segmentation of the treated (e.g. coiled) aneurysms will not be considered when assessing the performance of untreated, unruptured aneurysm detection or segmentation. That is: any false positive detections at the location of treated aneurysms will be ignored during evaluation.

For both tasks, the description of desired output will be provided to all participants with a full in-depth explanation and examples provided on the challenge website. Furthermore, the aneurysm mask, centre of mass and radius are provided for each training case so the participant can self-evaluate based on this. The source code that will be used for the final evaluation will also be provided to participants.

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).

A variety of Philips MRI scanners (1.5 and 3T) were used to acquire the TOF-MRA and T1/T2 images. An in-house developed software was used to develop the labels of the aneurysms from the TOF-MRA images.

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

Due to the clinical nature of this data set and that it was taken from multiple studies across many years, there is no set protocol or acquisition used across cases. This provides a diverse and realistic data set in which standard clinical protocols and routine were used.

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.

University Medical Center (UMC) Utrecht, The Netherlands

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).

Not applicable

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases both represent a Time of Flight MRA of a human brain and a corresponding structural image (either a T1 or T2 weighted MRI). Training and test cases will both have corresponding annotations of the aneurysm based on the TOF-MRA to use for training or evaluation of the method.
The desired result is a binary segmentation mask of the detected aneurysm in the TOF-MRA.

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

Each case consists of one TOF-MRA and one structural image.

Train Dataset: 110 cases
1. 90 cases containing at least one untreated, unruptured intracranial aneurysm
-> 35 baseline and 35 follow-up of the same subject
-> 20 unique subjects
2. 20 scans of subjects without intracranial aneurysms

Test Dataset (not released): 140 cases
1. 115 cases containing at least one untreated, unruptured intracranial aneurysm
-> 45 baseline and 45 follow-up of the same subject
-> 25 unique subjects
2. 25 cases of subjects without intracranial aneurysms

We are not providing a specific validation set and it is up to the participants to decide a validation/train set split based on the provided training data. Follow-up scans were performed at least 6 months after the baseline scan.

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

The number of training data was chosen as this allowed enough variability between scans and subjects to give a broad, diverse dataset representing a dataset in a real clinical setting. Previous studies based on similar detection/segmentation methods suggest that approximately 100 scans of subjects containing aneurysms would

be an adequate amount to allow detection/segmentation to be performed [7,13,14]. The size of the test data set allows the methods to be tested on a data set similar to that of a real-life dataset from the clinic. The test set also allows us, as the organising team, to perform detailed analysis on the different submitted methods and the variability of results.

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 scans used were real clinical data taken from a long time period 2001-2019. The structural image provided for each scan is either a T1 or T2 weighted MRI, depending on the protocol used. This means it includes a variety of different scanners, protocol parameters, resolutions and qualities. This represents a realistic clinical dataset, which makes the challenge more applicable and applicable to everyday clinical use.

The subjects with two scans of both baseline and follow-up represents real world data that would come from the clinic. It is normal practise that a patient with an unruptured aneurysm, considered not at risk of rupture, to undergo follow-up imaging some time later. It also allows the organisers to perform evaluation on how the algorithm works within scans of the same subject.

The training data will include scans with a distribution of different scan protocol parameters and different aneurysm characteristics. We will ensure that the training data is representative of the test data, by controlling the split according to important parameters such as aneurysm location and size, scanning protocol parameters and the age/sex of the patient. No subjects are in both test and training set. This will ensure that all types of aneurysms and images are present in both the training and test data sets, and that the training set is representative of the test set.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Manual segmentations were made on axial slices of the TOF-MRAs using in-house developed software. These were then converted to give a 3D binary mask of the aneurysm. The annotations were performed by two raters. Both raters were given training on the annotation software by the software developers.

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 protocol for aneurysm annotation was developed by an interventional neuro-radiologist. On each axial slice of the TOF-MRA a contour around the outline of the aneurysm was drawn. The annotations were always drawn to be from the level of the neck to the dome of the aneurysm. The neck corresponds to the opening of the aneurysm from the parent vessel. The dome is the furthest part of the aneurysm from the parent vessel. None of the parent vessel was included in the annotation. During annotation, the raters had access to the structural image and a radiologist report made at the time of the scan. The radiologists' reports indicated the rough location and size of the aneurysm.

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.

Aneurysms were diagnosed in all images as part of clinical routine. Upon inclusion of subjects in the ELAPSS/FAMSAB/ERASE study, an experienced interventional neuro-radiologist with more than 10 years of experience in the field manually detected each aneurysm and provided a detailed description and location. This neuro-radiologist trained a second rater with extensive experience in medical image analysis and the annotation software, but not specifically intracranial aneurysms. Once this second rater was on par with the first rater (as assessed by comparing intra/inter rater variability), the rater annotated all images in the dataset. Finally, the first and second rater will assess the full dataset together and make any required modifications in consensus. Hence, this will result in a consensus annotation that will be the official ground truth data set. The subset of annotations that have been annotated independently by both raters will be used to determine interrater variability.

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 multiple annotations will only be used to determine interrater variability. The actual annotations used will be those decided in consensus as described above.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The data will be pre-processed using n4 bias field correction [17], the face will be masked, and the structural image will be co-registered with the TOF-MRA. Both the original anonymised and the pre-processed data will be provided, alongside the face-mask used. This allows participants to implement their own pre-processing if they desire. A structural image (either T1 or T2) has been provided for each TOF-MRA and participants are welcome to use these as desired whether for pre-processing for inclusion in the actual training data. For the structural images, both the original image and a co-registered image will be provided, including the transformation parameters. The test data will not be provided to participants, but will undergo the exact same pre-processing as for the training data.

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.

Current findings, based on manual annotations made on a subset of the data-set by an experienced rater and a student, suggest that the manual annotations may vary by approximately 5% in volume. We have obtained a grant from Focus area Applied Data Science, Utrecht University to fund a second rater who will be trained by the experienced neuro-radiologist. Thus determination of the true magnitude of these errors will be in the spring of 2020 and be released on the challenge website. Based on a previous study using a similar manual segmentation technique [18] good inter- and intra-observer reliability of the volume was found with an intra-class correlation coefficient of 0.94.

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

The quality of the scans influences the variability of annotations between scans. However, the diverse range of

scans with different qualities and resolutions gives a real representation of a true clinical dataset. The size of the aneurysm also influences the annotation variability, as smaller aneurysms are understandably harder to annotate and voxel size has more of an influence.

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)

Segmentation:
- Dice Similarity Coefficient
- Hausdorff distance (modified, 95th percentile)
- Volumetric Similarity

Indication of how this metrics can be determined can be found here: [19]

These metrics will be determined for each scan in the test set. For the final evaluation, a mean of each of the metrics over all of the test scans will be determined. For all metrics, other than false positive count, the mean of the metrics will be weighted to the number of aneurysms per scan. This allows for evaluation of the performance of the method across all of the scans and aneurysms which is important due to the variance between scans.

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

Segmentation:
The chosen metrics are similar to those used in previous similar segmentation challenges [21] and are well used evaluation metrics in medical image segmentation.

Dice Similarity Coefficient – The Dice similarity coefficient is the most common metric used for evaluating medical image segmentations. It allows a clear understanding of the similarity of the segmentation versus the ground truth based on how much the two segmentations overlap.

Modified Hausdorff distance – The Hausdorff distance (in mm) was chosen as a metric as this allows the assessment of the accuracy of the boundary of the segmentation relative to the ground truth. This measure allows the actual spatial location of the voxels to be considered (i.e. the location of the aneurysm) as well as the size. As the Hausdorff distance is sensitive to outliers, the modified version was used where the maximum distance is defined as the 95th percentile. This is a good metric for small regions (such as small aneurysms) as they are more likely to have a small overlap and a smaller Dice. The distance metric is also more sensitive to shape of the segmentation, due to its reliance on the boundary of the shape. This is important when segmenting aneurysms as the shape may be used to assess rupture risk.

Volumetric similarity – The volumetric similarity compares the actual volumes of the segmentation and the ground truth. As the aneurysm size is important here, for example: for assessing aneurysm growth, the volumetric similarity is important. If the difference between the two images is less than or equal to the volumetric similarity

then it becomes more difficult to reliably detect growth.

It is worth noting that the detection/segmentation of the treated (e.g. coiled) aneurysms will not be considered when assessing the performance of untreated aneurysm detection or segmentation. Any false positive detections at the location of treated aneurysms will be ignored during evaluation. The above metrics will only be determined for untreated, unruptured aneurysms.

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.

Each metric, defined in item 26 a), is averaged over all test scans for each task. For each metric, the participating teams are sorted from best to worst. The best team receives a rank of 0 and the worst team a rank of 1; all other teams are ranked (0 -1) relative to their performance within the range of that metric. Finally, the ranks for each metric are averaged into the overall rank that is used for the results.

For example: the best team A has a DSC of 80 and the worst team B a DSC of 60. In the ranking: A=0.00 and B=1.00. Another team C has a DSC of 78, which is then ranked at 1.0 - (78 - 60) / (80 - 60) = 0.10.

This is the same ranking system used the WMH challenge, which worked well. It allows a fair and direct comparison of the evaluation of the different methods.

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

Not Applicable. No submissions will have missing results as we are performing the evaluation on test sets in house, using the code provided with Docker containers. If algorithm does not produce any output it will be considered as "no aneurysms detected". If we notice that a method has technical errors that prevent it from running we will contact the participants to have this fixed.

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

The described ranking scheme will be used to give equal weighting to all metrics (when they may have different ranges) and allow a fair comparison between teams. This has worked well for the WMH and MRBrainS challenges. The relative ranking clearly shows the small/large performance differences between competing methods; which does not show very well in conventional ranking (e.g. methods ranked 1 and 10 may be closer in performance than methods 13 and 16).

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.

No data will be missing as the test data will not be provided to participants and all evaluation is done at the organisers' institute. Evaluation will be performed using the metrics and ranking as described in the previous

section (26). The evaluation and ranking code will be published online beforehand.
The teams will be ranked from best to worst on an average of the ranking of all the metrics. 95% confidence intervals on each individual metric and on the final ranking will be computed using bootstrapping. This will be performed by randomly taking a large number of samples from the test set using sampling by replacement. Non-overlapping confidence intervals, determined in this way, indicate a significant difference between the methods.

Full statistical analysis of all methods will be detailed in the final journal paper after the conference.

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

95% confidence intervals and bootstrapping were used to ensure that the methods were statistically different from each other with regard to each metric.

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.

In further analyses, algorithms will be combined using an ensemble method in the journal paper.For the segmentation task, an algorithm such as the Simultaneous Truth And Performance Level Estimation (STAPLE) may be used on all methods. It takes multiple methods as an input and produces a combined segmentation which can be evaluated and ranked separately. Multiple methods to assess using STAPLE will be chosen based on the distribution and variability of the methods submitted. [22]

A ranking will also be produced based on the intra-subject performance to determine those methods that perform well to assess growth. This will be defined by investigating if the methods are capable of identifying growth, where the growth is a change of volume of the segmented aneurysm which exceeds the volumetric similarity. This will be compared to growth as it is clinically defined and differences investigated.

As both a TOF-MRA and a structural image (T1 or T2) are provided, analysis will also be performed on the modality of image used for each of the methods and how this affected the results and ranking. Analysis of the submitted algorithm performance compared to the interrater reproducibility will also be measured.

Analyses will also be performed on the different types and structures of methods used, however this analysis will be dependent on the type of methods submitted (e.g. machine-learning vs classical, different deep-learning techniques etc.) This may be performed in a similar way as in the WMH challenge paper. [21]

Full statistical analysis of all methods will be detailed in the final journal paper after the conference.

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] A. Keedy, “An overview of intracranial aneurysms,” McGill Journal of Medicine, vol. 9, no. 2. pp. 141–146, 2006.

[2] J. M. Wardlaw and P. M. White, "The detection and management of unruptured intracranial aneurysms," Brain, vol. 123, no. 2, pp. 205–221, 2000.
[3] P. M. White, J. M. Wardlaw, and V. Easton, "Can noninvasive imaging accurately depict intracranial aneurysms? A systematic review," Radiology, vol. 217, no. 2, pp. 361–370, 2000.
[4] P. M. White, E. M. Teasdale, J. M. Wardlaw, and V. Easton, "Intracranial aneurysms: CT angiography and MR angiography for detection - Prospective blinded comparison in a large patient cohort," Radiology, vol. 219, no. 3, pp. 739–749, 2001.
[5] M. J. H. Wermer, I. C. van der Schaaf, A. Algra, and G. J. E. Rinkel, "Risk of rupture of unruptured intracranial aneurysms in relation to patient and aneurysm characteristics: an updated meta-analysis.," Stroke, vol. 38, no. 4, pp. 1404–10, 2007.
[6] D. Backes et al., "ELAPSS score for prediction of risk of growth of unruptured intracranial aneurysms," Neurology, vol. 88, no. 17, pp. 1600–1606, 2017.
[7] A. E. Lindgren et al., "Irregular Shape of Intracranial Aneurysm Indicates Rupture Risk Irrespective of Size in a Population-Based Cohort," Stroke, vol. 47, no. 5, pp. 1219–1226, 2016.
[8] M. L. Raghavan, B. Ma, and R. E. Harbaugh, "Quantified aneurysm shape and rupture risk," J. Neurosurg., vol. 102, no. 2, pp. 355–362, 2009.
[9] W. Ji et al., "Risk score for neurological complications after endovascular treatment of unruptured intracranial aneurysms," Stroke, vol. 47, no. 4, pp. 971–978, 2016.
[10] C. M. Hentschke, O. Beuing, R. Nickl, and K. D. Tönnies, "Automatic cerebral aneurysm detection in multimodal angiographic images," IEEE Nucl. Sci. Symp. Conf. Rec., no. October, pp. 3116–3120, 2012.
[11] H. Arimura et al., "Computerized detection of intracranial aneurysms for three-dimensional MR angiography: Feature extraction of small protrusions based on a shape-based difference image technique," Med. Phys., vol. 33, no. 2, pp. 394–401, 2006.
[12] H. Duan, Y. Huang, L. Liu, H. Dai, L. Chen, and L. Zhou, "Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks," Biomed. Eng. Online, vol. 18, no. 1, p. 110, 2019.
[13] N. Sulayman, M. Al-Mawaldi, and Q. Kanafani, "Semi-automatic detection and segmentation algorithm of saccular aneurysms in 2D cerebral DSA images," Egypt. J. Radiol. Nucl. Med., 2016.
[14] A. Lane, P. Vivian, and A. Coulthard, "Magnetic resonance angiography or digital subtraction catheter angiography for follow-up of coiled aneurysms: Do we need both?," J. Med. Imaging Radiat. Oncol., vol. 59, no. 2, pp. 163–169, 2015.
[15] X. Yang, D. J. Blezek, L. T. E. Cheng, W. J. Ryan, D. F. Kallmes, and B. J. Erickson, "Computer-aided detection of intracranial aneurysms in MR angiography," J. Digit. Imaging, vol. 24, no. 1, pp. 86–95, 2011.
[16] A. Faron, R. Sijben, N. Teichert, J. Freiherr, M. Wiesmann, and T. Sichtermann, "Deep learning-based detection of intracranial aneurysms in 3D TOF-MRA," Am. J. Neuroradiol., vol. 40, no. 1, pp. 25–32, 2019.
[17] N. J. Tustison, P. A. Cook, and J. C. Gee, "N4Itk," vol. 29, no. 6, pp. 1310–1320, 2011.
[18] E. E. de Vries et al., "Volumetric assessment of extracranial carotid artery aneurysms," Sci. Rep., vol. 9, no. 1, p. 8108, 2019.
[19] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," BMC Med. Imaging, vol. 15, no. 1, 2015.
[20] B. van Ginneken et al., "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study," Med. Image Anal., 2010.
[21] H. J. Kuijf et al., "Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge," IEEE Trans. Med. Imaging, vol. 38, no. 11, pp. 2556–2568, 2019.
[22] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An

algorithm for the validation of image segmentation," IEEE Trans. Med. Imaging, vol. 23, no. 7, pp. 903–921, 2004.