VAscular Lesions DetectiOn - New: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

VAscular Lesions DetectiOn - New

Challenge acronym

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

Where is VALDO

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.

Appropriate blood supply is essential to the healthy maintenance of brain tissue. With age, vascular changes are observed in the smallest vessels resulting in impaired function. Changes to the surrounding tissue can be observed using magnetic resonance imaging. White matter hyperintensities (WMH) are one such prominent marker of cerebral small vessel disease (CSVD) and their automated segmentation has been the focus of a large body of research as well as of segmentation challenges. Other markers of CSVD exist and their quantification along with WMH is essential to grasp the overall picture of the vascular burden related to CSVD. They include notably lacunes, enlarged perivascular spaces and cerebral microbleeds. Manual annotations are extremely time-consuming and suffer greatly from inter- and intra-rater variability, due to their small size and the difficulty of distinguishing these markers from each other and similarly appearing structures as well as the lack of a way to uncover the "real" ground truth. However, many studies have hinted at their potential to become essential biomarkers. Automated methods are therefore required to make their quantification not only robust and reliable, but simply feasible. So far development of such methods has been impeded by the methodological issues related to their very small size and the extreme imbalance in the data, but also the absence of sufficient gold standard.

This challenge aims at promoting the development of new solutions for the automated segmentation of such very sparse and small objects while leveraging weak and noisy labels. The central objective of this challenge is to facilitate quantification of CSVD in brain MRI scans.

The challenge will have a technical impact in the following fields: object detection, segmentation, class imbalance, use of weak labels, multi-scale object detection, assessment of prediction uncertainty. The biomedical impact will not only directly impact the field of cerebral small vessel disease research but also other brain pathologies such as multiple sclerosis where similar objects have recently been shown renewed interest. More broadly, translation of developed techniques to other fields where sparse object detection is essential will be impacted.

Challenge keywords

List the primary keywords that characterize the challenge.

Extremely small objects - segmentation - detection - noisy labels - multi-task learning - brain - MRI - cerebral small vessel disease - vascular - perivascular spaces - microbleeds - lacunes

Year

The challenge will take place in ...

2021

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

None

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect at least 15 participants. We sent out a survey to assess the interest in the community, which indicated 12 people were very likely to participate in our challenge and 12 people might participate.
With this challenge we would like to encourage more people to work on this application and increase awareness of the interesting challenges associated to it. We would like to host this challenge in 2021, in order to appropriately advertise its existence and provide training data early enough for people new to the field to appropriately develop their methodological solutions. Additionally, this would provide us with enough time to prepare the data and annotations.

Publication and future plans

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

The challenge results and introspection on the proposed methods and outcomes will be then gathered for publication.

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 hosted on the Grand Challenge platform. Evaluation of the submissions will be done at the Erasmus Medical Center as part of the test set data cannot leave this center.

On the day of the challenge, a projector, a computer and two microphones will be needed in order to let the

participants describe their proposed solution and for the outcomes of the challenge to be announced.

TASK: Segmentation of enlarged PVS

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

The burden of enlarged PVS is currently emerging as an important neuroimaging biomarker. The current bottleneck for studying PVS burden is the need for an automated method. Manual annotation of enlarged PVS is extremely time-consuming due to the large number of enlarged PVS that can be present in MRI scans. Currently, neurological studies mostly score the burden of enlarged PVS visually by e.g. counting the number of enlarged PVS in a slice. This is the most practical and fast way to quantify enlarged PVS, however it is a coarse approximation of the large amount of valuable information in the scans. Furthermore, manual annotation and visual scoring are subject to observer bias due to the difficulty of assessing if a PVS is enlarged and distinguishing it from other similarly appearing structures. Dealing with this subjectivity of annotations is one of the main challenges for current automated methods, as it is not possible to acquire a "real" ground truth.
A robust, automated method for segmenting enlarged PVS would be extremely useful for neurological research on the role of enlarged PVS in neurological disorders.

Keywords

List the primary keywords that characterize the task.

Enlarged PVS - segmentation - detection - noisy labels - weak label - count - visual score - extremely small

ORGANIZATION

Organizers

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

Kimberlin van Wijnen, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands
Carole Sudre, School of Biomedical Engineering & Imaging Sciences - King's College London - London - United Kingdom / Dementia Research Centre University College London - London - United Kingdom
Marius de Groot, GSK - London - United Kingdom
Florian Dubost, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands
Marleen de Bruijne, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands, Department of Computer Science - University of Copenhagen - Copenhagen - Denmark

b) Provide information on the primary contact person.

Kimberlin van Wijnen, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Open call challenge.

Challenge venue and platform

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

MICCAI.

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

Grand-challenge.org.

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

https://valdo.grand-challenge.org

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.

Publicly available data can be used for training as well as all challenge data, also those available for other tasks. Participants have to indicate whether they used additional private training data. This information will be displayed on the leaderboard, but will not be taken into account for the awards.

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.

Members of the organizers' primary institutes may participate in the challenge. Their submissions will be listed on the leaderboard with the information that they are from one of the organizers' primary institutes. They will not be eligible for awards.

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

Sponsors for the awards are currently being investigated. We aim to have an award for each winner of each task and for an overall winner (across all tasks).

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.

The top 3 methods will be announced publicly during MICCAI 2021. Submitted results have to contain a pdf describing the method. All submitted results will be evaluated and their ranks on multiple metrics published on the leaderboards that we will open up after the results have been presented at MICCAI 2021.

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).

From the participating teams, two authors will qualify as author for the resulting publication. Participating teams are free to publish their own results whenever they want. Any publication showing results on the training data should also show the results on challenge test data as obtained from submission to the challenge.

Submission method

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

Examples:

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

Participants have to submit their algorithms in the form of a Docker container, which will be applied to the test set by the organizing team to evaluate the performance on the test set. We will provide submission instructions on the challenge website as well as a Docker template.

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.

Participants will be allowed to submit their Docker container to be run on a small disjoint validation set from one center before their final submission. This set does not overlap with the training set nor with the test set. This will be allowed twice at most. Their final submission will be evaluated on the test set and will be officially counted for the challenge results.

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

February 2021: Release of training data
Mid April 2021 - Mid July 2021: Optional submission for evaluation on a small validation set from one center (see section pre-evaluation, part b)
Early August 2021: Final Docker container submission
Results on test set run by organisation team
MICCAI 2021: Results announced
After MICCAI 2021: Open public leaderboard and provide links to submitted Docker containers on challenge website (see section code availability, part c)

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).

Retrospective studies with existing local ethical approval.
SABRE: National Research Ethics Service Committee, London–Fulham (14/LO/0108)
RSS: Medical Ethics Committee of Erasmus University

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

The evaluation code will be made openly available on github.

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

Links to the Docker containers released by the participants who have given permission for this will be made available on the website. Participants will be encouraged to give permission for this and to make their code publicly available.

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.

Grants of organizers: NWO P15-26 (partly funded by Quantib), ZonMw 104003005, Alzheimer's Society (AS-JF-17-011)
Only the organizers will have access to the test case labels.
We are currently looking into which companies are willing to sponsor the challenge in the form of awards. Icometrix will sponsor the challenge with one award. We are also in contact with other potential sponsors.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

CAD, Prognosis, Research, Diagnosis, Screening.

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 an ageing population with vascular damage, cardiovascular risk factors and cognitive decline.

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

Cohort 1 (SABRE) is a tri-ethnic ageing population with high cardiovascular risk factors.
Cohort 2 (RSS) is a population study of an ageing population in a homogeneous environment.

Imaging modality(ies)

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

T1 weighted (T1w), T2 weighted (T2w), Fluid-attenuated inversion recovery (FLAIR) MRI scans

Context information

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

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

No additional data will be given along with the images.

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

No additional information will be given along with the images.

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.

Brain MRI scan - T1w, T2w, FLAIR

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

The structures that the participating algorithms should focus on are enlarged perivascular spaces (PVS).

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)).

Precision, Specificity, Accuracy, Reliability, Consistency, Sensitivity.

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).

Cohort 1, Southall and Brent Revisited (SABRE): 3T MRI scanner Philips
Cohort 2, Rotterdam Scan Study (RSS): 1.5T MRI scanner General Electric (GE) Healthcare

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

Cohort 1, SABRE:
T1w: inversion-prepared gradient echo, TR/TE = 6.9/3.1 ms, voxel size = 1.09 x 1.09 x 1.0 mm3.
FLAIR: TR/TE/TI = 4800/125/1650 ms, voxel size = 1.09 x 1.09 x 1.0 mm3
T2w 3D: sagittal, turbo spin echo, TR/TE/TI = 2500/222 ms, voxel size = 1.09 x 1.09 x 1.09 mm3.

Cohort 2, RSS:
FLAIR: fast spin echo, TR/TE/TI = 8000/120/2000 ms, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.
T1w: gradient-recalled echo, TR/TE/TI 13.8/2.8/400 ms, flip angle = 20°, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.
T2w: fast spin echo, TR/TE = 12300/17.3 ms, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.

Repetition time (TR), echo time (TE), inversion time (TI), T1 weighted (T1w), T2 weighted (T2w), Fluid-attenuated inversion recovery (FLAIR).

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.

Cohort 1, SABRE: University College London
Cohort 2, RSS: Erasmus Medical Center

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).

Trained radiographers acquired the data according to a predefined research protocol.

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.

Both training and test cases represent coregistered T1w, T2w and FLAIR MRI scans (registered to the T1w MRI scan) of a human brain. The RSS training cases contain 1 set of annotations - counts (for techniques using also weak labels) or segmentations - either in specified brain regions or in whole slices. The SABRE training cases contain 2 sets of segmentations (two raters) in slabs (stacks of slices). Every case also contains a mask indicating for which parts of the brain annotations are provided. Test cases contain segmentations in part of the brain or in the whole brain.

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

Train: 40 cases (6 cases from SABRE, 34 cases from RSS), representing in total 40 subjects.
Test: 66 cases (10 cases from SABRE, 56 cases from RSS, same ratio as in training set), representing in total 66 subjects.

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

The number of cases was decided based on the limited number of available scans with segmentations and the number of cases we could make publicly available. The ratio between training and test set was applied similarly to all centers providing data and raters providing annotations (approx. 40% train, 60% test).

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 training and test cases were chosen to represent the variability of number of enlarged PVS and subject age in the two studies. We applied stratified random sampling aimed at acquiring an approximately uniform distribution over the number of enlarged PVS and the age of the subjects. The distribution of number of enlarged PVS is similar in training and test sets.

The training set subjects of SABRE overlap with 6 of the cases in the microbleed task and with all of the subjects in the lacune task. The RSS subjects of the training set are the same for all tasks.

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.

Train:
6 cases from SABRE with manual segmentations of rater 1 and rater 2 in slabs (=stacks of slices, for tridimensional information).
6 cases from RSS with manual segmentations from rater 3 in two slices per region or in full regions*.
28 cases from RSS with counts from rater 4 in one slice per region or in full regions*.
The chosen annotated slices are distributed throughout the brain, the aim is to include all brain morphology.

Test:
10 cases from SABRE with manual segmentations of rater 1 and rater 2 in the full brain.
56 cases from RSS with manual segmentations in 2 slices per region or in full regions*. 44 cases were segmented by rater 3 and 12 cases were segmented by rater 3 and 5.

* For the two larger regions (the centrum semiovale (CSO) and the basal ganglia (BG)) annotations are provided in 1 or 2 slices. For the two smaller regions (hippocampi and midbrain) annotations in the full regions are provided.

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.

Both: STRIVE criteria (see reference section) - Emphasis was given on providing 3D consistent segmentation and on looking at the three modalities simultaneously.

RSS: UNIVRSE criteria (see reference section) - Only enlarged perivascular spaces between 1 mm and 3 mm are considered. Raters were instructed to consider the three modalities.

For SABRE the STRIVE criteria (see reference section) were used:
- "Fluid-filled spaces that follow the typical course of a vessel as it goes through grey or white matter. The spaces have signal intensity similar to CSF on all sequences." [6]
- "Because they follow the course of penetrating vessels, they appear linear when imaged parallel to the course of the vessel, and round or ovoid, with a diameter generally smaller than 3 mm, when imaged perpendicular to the course of the vessel." [6]

For the RSS cases the following protocol was used:
- Enlarged PVS were defined using the UNIVRSE rating system [7], which is very similar to the STRIVE criteria. However only enlarged perivascular spaces larger than 1 mm are annotated (in the STRIVE criteria there is no lower threshold for size, only the upper threshold of 3 mm)
- Enlarged PVS were annotated in four brain regions: the centrum semiovale (CSO), basal ganglia (BG), hippocampi and the midbrain [7].
- For the two larger brain regions (CSO and BG), a specific slice was annotated as described in the UNIVRSE rating system: for the CSO "the slice 1 cm above the lateral ventricles" [7] was annotated; for the basal ganglia "the slice showing the anterior commissure or, when not visible, the first slice superior to it." [7]. Additionally, one slice per region was randomly selected and annotated.
- For the two smaller regions, the hippocampi and the midbrain, the full regions were annotated, as in the UNIVRSE rating system [7].

For the segmentations, the raters annotated using contours that were subsequently converted to voxelwise labels. The counts were computed from manual dot annotations.

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.

SABRE: The segmentations were done by one medically trained rater (rater 1) and one rater with 5+ years of professional experience (rater 2). Software based correction was applied on the segmentation that verified that segmentations had appropriate signal intensities in the various MRI sequences (hypointense on T1w and FLAIR scans, hyperintense on T2w scans).

RSS:
The segmentations were done by medical students with annotation experience that were trained to recognize enlarged perivascular spaces (rater 3 and 5).
The counts were annotated by a medical student with annotating experience (rater 4).

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.

For SABRE cases annotations of two raters will be made available.

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.

Rigid coregistration of T1w, T2w and FLAIR MRI scans to the T1w MRI scan. Per case a mask is available indicating for which parts of the brain annotations are provided. Specific brain regions (see description of annotations) will be provided in the mask, so these brain regions can be extracted from the MRI scans if necessary.

All RSS MRI scans were nonuniformity corrected (NU) with the MINC N3 package

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.

Possible sources of error with respect to annotating enlarged PVS in general are the difficulty of assessing if PVS are enlarged enough, distinguishing enlarged PVS from similarly appearing structures, and in general the lack of a way to uncover the "real" ground truth. Additionally, enlarged PVS can be very small and easy to miss.

Possible sources of error in the segmentation pertain both to the identification by the operator of the appropriate element and to the definition of elements borders. Furthermore, deciding until where an enlarged PVS is still visible (on which slice or where in a slice) is difficult and can lead to errors. Further inacurracies may be due to issues in the use of the segmentation software tool (too large brush, not considering all orientations...) as well as initial misalignment. For the RSS cases with segmentations a possible error source is the conversion from contours (done by the raters) to voxelwise annotations (segmentation mask).

Possible sources of error specifically for counts relate to the choice of the slice on which to perform the assessment as well as possible observational shift.

Inter-rater variability will be assessed on a subset of the test cases.
(As of date - the second rater is still creating the segmentation on the SABRE data - inter rater variability will be evaluated once these annotations are done)

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

Another source of error may come from the rigid registration of the scans to the T1w MRI scan.

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)

We will compute the following metrics for cases with manual segmentations (averaged over non-empty cases):

Segmentation: Average Dice Similarity Coefficient (DSC) over detected elements
Volume: Absolute Difference
Detection: Detection F1
Count: Absolute Difference

Defined as:
DSC (voxelwise), Detection F1 (elementwise) = 2TP / (2TP + FN + FP)
Absolute Difference = |Total Predicted - Total Rater| = |FP - FN|

We will compute the following metrics for cases with no manual segmentations, so for all empty cases (averaged over empty cases):

Volume: Absolute difference
Count: Absolute difference

Defined as:
Absolute Difference = |Total Predicted - Total Rater| = |Total Predicted| = |FP|

All metrics will be computed based on 1 prediction threshold, namely at 0.5. Detection will be computed based on maximum overlap. Detection and count metrics and elementwise volume correlation will be averaged over 3 connectivity options (6, 18 and 26).

For test cases with multiple raters, volume and counts metrics will be averaged over raters. Segmentation and detection metrics will be weighed with the raters agreement:
Agreement: voxel/element has weight = 1
Disagreement: voxel/element has weight = 0.5
Case has only 1 rater: voxel/element has weight = 1

All metrics will be used to compute the ranking for this task, as well as to compute the ranking over all 3 tasks.

TP: true positives, FP: false positives, FN: false negatives

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

Submitted methods are expected to output a soft segmentation. We apply a threshold of 0.5 on these predicted soft segmentations. We compute the defined metrics for all test cases and then average over these cases.

In the evaluation of this task we focus on four important aspects of enlarged PVS annotations, namely segmentation, volume, detection and count.

Segmentation:
There is an extreme imbalance between background and foreground voxels for enlarged PVS segmentation. For this reason we compute DSC to evaluate segmentation performance, this is a commonly used metric that handles imbalance well. DSC is defined as two times the intersection between the prediction and manual segmentations (the number of voxels that are correctly predicted as enlarged PVS, true positives) divided by the total number of positive voxels (enlarged PVS voxels in manual segmentations) and the total number of voxels predicted to be positive (enlarged PVS voxels in prediction). It varies between 0 (no enlarged PVS voxels correct) and 1 (all enlarged PVS voxels are correctly predicted and no missed or wrong voxels).

Detection:

We will evaluate the detection performance using the commonly used Detection F1. This metric is equivalent to DSC, but is computed at the element level while DSC is computed at the voxel level. Elements are true positive when the intersection over union (IOU) is above the chosen threshold. As enlarged PVS can occur close together and in large amounts, an IOU threshold is more reliable for this task than a distance threshold (as is used in the other tasks). The IOU threshold will be chosen by comparing the segmentation masks of 2 raters (to be announced). This metric varies between 0 (no enlarged PVS detected correctly) and 1 (all enlarged PVS are correctly detected and no missed or wrong detections).

Volume:
We will evaluate the total predicted volume using the absolute difference between the predicted total volume and the rater's total segmented volume. The total (predicted) volume per case is the sum over all foreground voxels in the (predicted) segmentation mask.

Count:
We will evaluate counts using absolute difference between the predicted counts and the rater's counts. Counts are computed by computing the number of connected components per case (see elements part below).

Elements (relevant for detection and count metrics):
Elements are extracted from the segmentation masks with connected component labeling. Choosing the connectivity is not straightforward as enlarged PVS are very small, but can be located very close together. As this could considerably affect the performance, we will compute the elements using a connectivity of 6, 18 and 26 and average the detection metric (detection F1) and count metric (absolute difference) over these 3 options.

Multiple raters (relevant for all metrics):
For part of the test set we have segmentations of two different raters.
- For the volume and counts metrics the metrics will be computed per rater and the metrics will be averaged over the raters.
- The segmentation (DSC) and detection (Detection F1) metrics are weighed by the raters agreement. For segmentation: voxels that the raters agree on have a weight of 1, voxels that raters disagree on have a weight of 0.5. For test cases with only 1 rater all voxels have a weight of 1. The weighed metrics quantify the similarity between the predictions and the segmentations weighed by how certain the raters are per voxel. The same weighing is applied for the detection metric, however per element instead of per voxel.

For empty cases (no enlarged PVS in scan):
When cases have no enlarged PVS (empty cases) there can be no TPs or FNs, so DSC and Detection F1 are not computed for these cases. The rater's count and volume are both 0 for these cases, so the absolute difference is equal to the total predicted amount, which in this case is the number of false positives.

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.

In order to achieve to final ranking, we will apply the ranking methods described in the Medical Decathlon challenge. For each given metric, and for each participant, the score attributed is the number of other participants performing significantly worse than the given participant. The rank of each participating method is calculated over

this score (the higher, the better). The sum of these ranks across metrics is then used as the overall performance, with the lowest rank corresponding to the best performing algorithm.
All sections of the evaluation (volume, count, segmentation and detection) count equally in the overall ranking. As there is only 1 metric per section for this task, all metrics are weighed equally. We will weigh metrics for non-empty cases and empty cases with the corresponding number of cases.

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

For submissions that result in missing outputs, all missing cases will be given the worst corresponding metrics value if the associated metrics is bounded and a 10% worse value than the worst observed metric across all other participants for the corresponding case if the metric is not bounded.

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

This is to date the most robust way of providing unbiased ranks when deciding on a collection of metrics.

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.

Wilcoxon signed-rank test.
Bootstrapping test set, computed confidence interval.

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

We will use bootstrapping to simulate the performance of methods for a new set drawn from the distribution. We will use the Wilcoxon signed-rank test to test significance as we will have paired data and the error distributions will probably not be normally distributed. The Wilcoxon signed-rank test is a nonparametric statistical test for paired data.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Further analyses will include inter algorithm variability in terms of the methods and clustering of results. Performance of ensembling methods based either on all or the top 50% will be assessed. Methods will be assessed by performance per center and vascular burden to assess any existing bias.

TASK: Segmentation of cerebral microbleeds

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Cerebral microbleeds are an essential marker of cerebral small vessel disease. Their presence has been associated with specific vascular pathology such as cerebral amyloid angiopathy and with other markers of cerebral small vessel disease (WMH, enlarged PVS). Currently, microbleeds are largely identified manually. The challenges of automated identification of microbleeds are the presence of numerous mimics and the sparsity of the data: microbleeds are very small and most often there are very few microbleeds per scan. An automated method for microbleed segmentation would enable further research on their presence in the context of neurodegenerative diseases.

Keywords

List the primary keywords that characterize the task.

Cerebral microbleeds - segmentation - detection - extremely small

ORGANIZATION

Organizers

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

Carole Sudre, School of Biomedical Engineering & Imaging Sciences - King's College London - London - United Kingdom / Dementia Research Centre University College London - London - United Kingdom
Kimberlin van Wijnen, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands
Marius de Groot, GSK - London - United Kingdom
Florian Dubost, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands
Marleen de Bruijne, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands, Department of Computer Science - University of Copenhagen - Copenhagen - Denmark

b) Provide information on the primary contact person.

Carole Sudre, School of Biomedical Engineering & Imaging Sciences - King's College London - London - United Kingdom / Dementia Research Centre University College London - London - United Kingdom

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Open call challenge.

Challenge venue and platform

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

MICCAI.

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

Grand-challenge.org.

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

https://valdo.grand-challenge.org

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.

Publicly available data can be used for training as well as all challenge data, also those available for other tasks. Participants have to indicate whether they used additional private training data. This information will be displayed on the leaderboard, but will not be taken into account for the awards.

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.

Members of the organizers' primary institutes may participate in the challenge. Their submissions will be listed on the leaderboard with the information that they are from one of the organizers' primary institutes. They will not be eligible for awards.

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

Sponsors for the awards are currently being investigated. We aim to have an award for each winner of each task and for an overall winner (across all tasks).

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.

The top 3 methods will be announced publicly during MICCAI 2021. Submitted results have to contain a pdf describing the method. All submitted results will be evaluated and their ranks on multiple metrics published on the leaderboards that we will open up after the results have been presented at MICCAI 2021.

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).

From the participating teams, two authors will qualify as author for the resulting publication. Participating teams are free to publish their own results whenever they want. Any publication showing results on the training data should also show the results on challenge test data as obtained from submission to the challenge.

Submission method

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

Examples:

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

Participants have to submit their algorithms in the form of a Docker container, which will be applied to the test set by the organizing team to evaluate the performance on the test set. We will provide submission instructions on the challenge website as well as a Docker template.

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.

Participants will be allowed to submit their Docker container to be run on a small disjoint validation set from one center before their final submission. This set does not overlap with the training set nor with the test set. This will be allowed twice at most. Their final submission will be evaluated on the test set and will be officially counted for the challenge results.

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

February 2021: Release of training data
Mid April 2021 - Mid July 2021: Optional submission for evaluation on a small validation set from one center (see section pre-evaluation, part b)
Early August 2021: Final Docker container submission
Results on test set run by organisation team
MICCAI 2021: Results announced
After MICCAI 2021: Open public leaderboard and provide links to submitted Docker containers on challenge website (see section code availability, part c)

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).

Retrospective studies with existing local ethical approval.
SABRE: National Research Ethics Service Committee, London–Fulham (14/LO/0108)
RSS: Medical Ethics Committee of Erasmus University,
ALFA: Independent Ethics Committee Parc de Salut Mar Barcelona and registered at Clinicaltrials.gov (Identifier: NCT01835717)

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

The evaluation code will be made openly available on github.

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

Links to the Docker containers released by the participants who have given permission for this will be made available on the website. Participants will be encouraged to give permission for this and to make their code publicly available.

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.

Grants of organizers: NWO P15-26 (partly funded by Quantib), ZonMw 104003005, Alzheimer's Society (AS-JF-17-011)
Only the organizers will have access to the test case labels.
We are currently looking into which companies are willing to sponsor the challenge in the form of awards. Icometrix will sponsor the challenge with one award. We are also in contact with other potential sponsors.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

CAD, Prognosis, Research, Diagnosis, Screening.

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 an ageing population with vascular damage, cardiovascular risk factors and cognitive decline.

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

Cohort 1 (SABRE) is a tri-ethnic ageing population with high cardiovascular risk factors.

Cohort 2 (RSS) is a population study of an ageing population in a homogeneous environment.
Cohort 3 (ALFA) contains cognitively normal participants with a family history of Alzheimer's disease (AD) and with a relatively high percentage of participants that are APOE-e4 carriers (main genetic risk factor for sporadic AD).

Imaging modality(ies)

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

T1 weighted (T1w), T2* weighted (T2*w), T2 weighted (T2w) MRI scans

Context information

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

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

No additional data will be given along with the images.

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

No additional information will be given along with the images.

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.

Brain MRI scan - T1w, T2*w, T2w

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

The structures that the participating algorithms should focus on are cerebral microbleeds.

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)).

Precision, Specificity, Accuracy, Reliability, Sensitivity.

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).

Cohort 1, Southall and Brent Revisited (SABRE): 3T MRI scanner Philips
Cohort 2, Rotterdam Scan Study (RSS): 1.5T MRI scanner General Electric (GE) Healthcare
Cohort 3, ALFA: 3T MRI scanner GE Discovery

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

Cohort 1, SABRE:
T1w: inversion-prepared gradient echo, TR/TE = 6.9/3.1 ms, voxel size = 1.09 x 1.09 x 1.0 mm3.
T2*w: gradient-echo,TR/TE = 1288/21, flip angle = 18°, voxel size reconstructed = 0.45 x 0.45 x 3.0 mm3.
T2w 3D: sagittal, turbo spin echo, TR/TE/TI = 2500/222 ms, voxel size = 1.09 x 1.09 x 1.09 mm3.

Cohort 2, RSS:
T1w: gradient-recalled echo, TR/TE/TI 13.8/2.8/400 ms, flip angle = 20°, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.
T2*w: gradient-recalled echo, TR/TE = 45/31 ms, flip angle = 13°, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.
T2w: fast spin echo, TR/TE = 12300/17.3 ms, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.

Cohort 3, ALFA:
T1w 3D: TR/TE/TI = 8.0/3.7/450 ms, flip-angle = 8°, voxel size = 1.0 x 1.0 x 1.0 mm3.
T2*w: gradient echo, TR/TE = 1300/23 ms, flip angle = 15°, voxel size = 1.0 x 1.0 x 3.0 mm3.
T2w: fast spin echo, TR/TE=5000/85ms, flip angle=110°, voxel size = 1.0 x 1.0 x 3.0 mm3

Repetition time (TR), echo time (TE), inversion time (TI), T1 weighted (T1w), T2* weighted (T2*w), T2 weighted (T2w)

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.

Cohort 1, SABRE: University College London
Cohort 2, RSS: Erasmus Medical Center
Cohort 3, ALFA: Barcelona Brain Research Center

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).

Trained radiographers acquired the data according to a predefined research protocol.

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.

Both training and test cases represent registered T1w, T2w and T2*w MRI scans (registered to the T2*w MRI scan) of a human brain. Both training and test cases contain one full annotation (as in segmentations) of cerebral microbleeds in the whole brain.

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

Train: 74 cases (10 cases from SABRE, 34 cases from RSS, 30 cases from ALFA), representing in total 74 subjects.
Test: 148 cases (20 cases from SABRE, 68 cases from RSS, 60 cases from ALFA, same ratio as in training set), representing in total 148 subjects.

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

The number of cases is limited by the effort of the labeling task and the number of cases we could make publicly available. The ratio between training and test set was applied similarly to all centers providing data and raters providing annotations (approx. 33% train, 67% test).

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 training and test cases were chosen to represent the variability in number of microbleeds and subject age. We applied stratified random sampling aimed at acquiring a uniform distribution over the number of microbleeds and the age of the subjects. The training set is a representative set of the test set.

The training set subjects of SABRE overlap with 6 of the subjects in the enlarged PVS task and the lacune task. The remaining 4 subjects do not overlap with the training nor test set of the other tasks. The RSS subjects of the training set are the same for all tasks. The subjects of the ALFA study do not overlap with the sets of the other tasks.

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.

All cases contain one set of segmentations.

Train:
10 scans from SABRE with segmentations in the full brain.
34 scans from RSS with segmentations in the full brain.
30 scans from ALFA with segmentations in the full brain.

Test:
20 scans from SABRE with segmentations in the full brain.
68 scans from RSS with segmentations in the full brain.
60 scans from ALFA with segmentations in the full brain.

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.

Microbleeds were segmented on T2*w MRI scans for all cases.

For the SABRE and ALFA cases the BOMBS criteria was used [8].

For the RSS cases the following guidelines were used:
- "Microbleeds were defined as focal areas of very low signal intensity, smaller than 10 mm in size." [9]
- This is in line also with the STRIVE criteria: "Small (generally 2–5 mm in diameter, but sometimes up to 10 mm) areas of signal void with associated blooming seen on T2*-weighted MRI or other sequences that are sensitive to susceptibility effects." [6]
Various different raters indicated the location of microbleeds on T2*w MRI scans. Based on these locations one rater segmented all microbleeds for all cases.

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.

SABRE: The segmentations were done by two human raters supervised by an expert neuroradiologist.
RSS: The locations of microbleeds were indicated by several expert raters (every case was looked at by one expert). Based on these locations one expert rater segmented all microbleeds.
ALFA: The segmentations were done by one medically trained rater (rater 1 from the enlarged PVS task) and one rater in training and were supervised by an expert neuroradiologist with 10 + years of professional experience.

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.

NA

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.

Rigid registration of T1w and T2w MRI scan to the T2*w MRI scan.

All RSS MRI scans were nonuniformity corrected (NU) with the MINC N3 package

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.

A possible source of error is the difficulty of identifying cerebral microbleeds and differentiating them from mimics. Another source lies in identifying the border of the element, due to partial volume effects it can be difficult to see which voxels are part of the microbleed and which are not. This is the case in axial slices, but also in depth, to see until which slice the microbleed is still visible.
Further inacurracies may be due to issues in the use of the segmentation software tool (too large brush, not considering all orientations...).

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

Another source of error may come from the rigid registration of the scans to the T2*w MRI scan.

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)

We will compute the following metrics for cases with manual segmentations (averaged over non-empty cases):

Segmentation: Average Dice Similarity Coefficient (DSC) over detected elements
Volume: Absolute Difference
Detection: Detection F1
Count: Absolute Difference

Defined as:
DSC (voxelwise), Detection F1 (elementwise) = 2TP / (2TP + FN + FP)
Absolute Difference = $|\text{Total Predicted} - \text{Total Rater}| = |\text{FP} - \text{FN}|$

We will compute the following metrics for cases with no manual segmentations, so for all empty cases (averaged over empty cases):

Volume: Absolute difference
Count: Absolute difference

Defined as:
Absolute Difference = $|\text{Total Predicted} - \text{Total Rater}| = |\text{Total Predicted}| = |\text{FP}|$

All metrics will be computed based on 1 prediction threshold, namely at 0.5. Detections will be evaluated using a distance threshold (distance between centers of mass).

All metrics will be used to compute the ranking for this task, as well as to compute the ranking over all 3 tasks.

TP: true positives, FP: false positives, FN: false negatives

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

Submitted methods are expected to output a soft segmentation. We apply a threshold of 0.5 on these predicted soft segmentations. We compute the defined metrics for all test cases and then average over these cases.

In the evaluation of this task we focus on four important aspects of microbleed annotations, namely segmentation, volume, detection and count.

Segmentation:
There is an extreme imbalance between background and foreground voxels for microbleed segmentation. For this reason we compute DSC to evaluate segmentation performance, this is a commonly used metric that handles imbalance well. DSC is defined as two times the intersection between the prediction and manual segmentations (the number of voxels that are correctly predicted as microbleed, true positives) divided by the total number of positive voxels (microbleed voxels in manual segmentations) and the total number of voxels predicted to be positive (microbleed voxels in prediction). It varies between 0 (no microbleed voxels correct) and 1 (all microbleed voxels are correctly predicted and no missed or wrong voxels).

Detection:
We will evaluate the detection performance using the commonly used Detection F1. This metric is equivalent to DSC, but is computed at the element level while DSC is computed at the voxel level. Elements are true positive when the distance between the centers of mass is below a chosen threshold (to be announced). This metric varies between 0 (no microbleeds detected correctly) and 1 (all microbleeds are correctly detected and no missed or wrong detections).

Volume:
We will evaluate the total predicted volume using the absolute difference between the predicted total volume and the rater's total segmented volume. The total (predicted) volume per case is the sum over all foreground voxels in the (predicted) segmentation mask.

Count:
We will evaluate counts using absolute difference between the predicted counts and the rater's counts. Counts are computed by computing the number of connected components per case (see elements part below).

Elements (relevant for detection and count metrics):
Elements are extracted from the segmentation masks with connected component labeling. We use a connectivity of 6 (only including direct neighbors).

For empty cases (no microbleeds in scan):
When cases have no microbleeds (empty cases) there can be no TPs or FNs, so DSC and Detection F1 are not computed for these cases. The rater's count and volume are both 0 for these cases, so the absolute difference is equal to the total predicted amount, which in this case is the number of false positives.

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.

In order to achieve to final ranking, we will apply the ranking methods described in the Medical Decathlon

challenge. For each given metric, and for each participant, the score attributed is the number of other participants performing significantly worse than the given participant. The rank of each participating method is calculated over this score (the higher, the better). The sum of these ranks across metrics is then used as the overall performance, with the lowest rank corresponding to the best performing algorithm.
All sections of the evaluation (volume, count, segmentation and detection) count equally in the overall ranking. As there is only 1 metric per section for this task, all metrics are weighed equally. We will weigh metrics for non-empty cases and empty cases with the corresponding number of cases.

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

For submissions that result in missing outputs, all missing cases will be given the worst corresponding metrics value if the associated metrics is bounded and a 10% worse value than the worst observed metric across all other participants for the corresponding case if the metric is not bounded.

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

This is to date the most robust way of providing unbiased ranks when deciding on a collection of metrics.

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.

Wilcoxon signed-rank test.
Bootstrapping test set, computed confidence interval.

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

We will use bootstrapping to simulate the performance of methods for a new set drawn from the distribution. We will use the Wilcoxon signed-rank test to test significance as we will have paired data and the error distributions will probably not be normally distributed. The Wilcoxon signed-rank test is a nonparametric statistical test for paired data.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Further analyses will include inter algorithm variability in terms of the methods and clustering of results. Performance of ensembling methods based either on all or the top 50% will be assessed. Methods will be assessed by performance per center and vascular burden to assess any existing bias.

TASK: Segmentation of lacunes

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.

Lacunes of presumed vascular origin are another important biomarker for cerebral small vessel disease. Currently, lacunes are generally detected manually, which is very time-consuming and subjective. Important challenges for automated methods are the small size of lacunes and their rare occurrence.
Lacunes can look very similar to enlarged PVS; they are often mistaken for each other. Even for experts distinguishing lacunes from enlarged PVS can be challenging and sometimes impossible depending on the MRI scan quality. This is more problematic for lacunes than for enlarged PVS, as the prevalence of lacunes is substantially lower than the prevalence of enlarged PVS. Automated methods should additionally focus on modeling the detection and segmentation uncertainty, as this is especially important for this biomarker.
An automated method for lacune segmentation would facilitate further research on lacunes and their role in neurological diseases.

Keywords

List the primary keywords that characterize the task.

Lacunes - segmentation - detection - noisy labels - uncertainty - extremely small

ORGANIZATION

Organizers

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

Carole Sudre, School of Biomedical Engineering & Imaging Sciences - King's College London - London - United Kingdom / Dementia Research Centre University College London - London - United Kingdom
Kimberlin van Wijnen, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands
Marius de Groot, GSK - London - United Kingdom
Florian Dubost, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands
Marleen de Bruijne, Biomedical Imaging Group - Erasmus Medical Center - Rotterdam - The Netherlands, Department of Computer Science - University of Copenhagen - Copenhagen - Denmark

b) Provide information on the primary contact person.

Carole Sudre, School of Biomedical Engineering & Imaging Sciences - King's College London - London - United Kingdom / Dementia Research Centre University College London - London - United Kingdom

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Open call challenge.

Challenge venue and platform

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

MICCAI.

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

Grand-challenge.org.

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

https://valdo.grand-challenge.org

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.

Publicly available data can be used for training as well as all challenge data, also those available for other tasks. Participants have to indicate whether they used additional private training data. This information will be displayed on the leaderboard, but will not be taken into account for the awards.

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.

Members of the organizers' primary institutes may participate in the challenge. Their submissions will be listed on the leaderboard with the information that they are from one of the organizers' primary institutes. They will not be eligible for awards.

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

Sponsors for the awards are currently being investigated. We aim to have an award for each winner of each task and for an overall winner (across all tasks).

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.

The top 3 methods will be announced publicly during MICCAI 2021. Submitted results have to contain a pdf describing the method. All submitted results will be evaluated and their ranks on multiple metrics published on the leaderboards that we will open up after the results have been presented at MICCAI 2021.

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).

From the participating teams, two authors will qualify as author for the resulting publication. Participating teams are free to publish their own results whenever they want. Any publication showing results on the training data should also show the results on challenge test data as obtained from submission to the challenge.

Submission method

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

Examples:

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

Participants have to submit their algorithms in the form of a Docker container, which will be applied to the test set by the organizing team to evaluate the performance on the test set. We will provide submission instructions on the challenge website as well as a Docker template.

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.

Participants will be allowed to submit their Docker container to be run on a small disjoint validation set from one center before their final submission. This set does not overlap with the training set nor with the test set. This will be allowed twice at most. Their final submission will be evaluated on the test set and will be officially counted for the challenge results.

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

February 2021: Release of training data
Mid April 2021 - Mid July 2021: Optional submission for evaluation on a small validation set from one center (see section pre-evaluation, part b)
Early August 2021: Final Docker container submission
Results on test set run by organisation team
MICCAI 2021: Results announced
After MICCAI 2021: Open public leaderboard and provide links to submitted Docker containers on challenge website (see section code availability, part c)

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).

Retrospective studies with existing local ethical approval:
SABRE: National Research Ethics Service Committee, London–Fulham (14/LO/0108)
RSS: Medical Ethics Committee of Erasmus University,

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

The evaluation code will be made openly available on github.

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

Links to the Docker containers released by the participants who have given permission for this will be made available on the website. Participants will be encouraged to give permission for this and to make their code publicly available.

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.

Grants of organizers: NWO P15-26 (partly funded by Quantib), ZonMw 104003005, Alzheimer's Society (AS-JF-17-011)
Only the organizers will have access to the test case labels.
We are currently looking into which companies are willing to sponsor the challenge in the form of awards. Icometrix will sponsor the challenge with one award. We are also in contact with other potential sponsors.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

CAD, Prognosis, Research, Diagnosis, Screening.

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 an ageing population with vascular damage, cardiovascular risk factors and cognitive decline.

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

Cohort 1 (SABRE) is a tri-ethnic ageing population with high cardiovascular risk factors.
Cohort 2 (RSS) is a population study of an ageing population in a homogeneous environment.

Imaging modality(ies)

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

T1 weighted (T1w), T2 weighted (T2w), Fluid-attenuated inversion recovery (FLAIR) MRI scans

Context information

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

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

No additional data will be given along with the images.

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

No additional information will be given along with the images.

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.

Brain MRI scan - T1w, T2w, FLAIR

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

The structures that the participating algorithms should focus on are lacunes of presumed vascular origin.

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)).

Precision, Specificity, Accuracy, Reliability, Sensitivity.

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).

Cohort 1, Southall and Brent Revisited (SABRE): 3T MRI scanner Philips
Cohort 2, Rotterdam Scan Study (RSS): 1.5T MRI scanner General Electric (GE) Healthcare

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

Cohort 1, SABRE:
T1w: inversion-prepared gradient echo, TR/TE = 6.9/3.1 ms, voxel size = 1.09 x 1.09 x 1.0 mm3.
FLAIR: TR/TE/TI = 4800/125/1650 ms, voxel size = 1.09 x 1.09 x 1.0 mm3.
T2w 3D: sagittal, turbo spin echo, TR/TE/TI = 2500/222 ms, voxel size = 1.09 x 1.09 x 1.09 mm3.

Cohort 2, RSS:
FLAIR: fast spin echo, TR/TE/TI = 8000/120/2000 ms, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.
T1w: gradient-recalled echo, TR/TE/TI 13.8/2.8/400 ms, flip angle = 20°, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.
T2w: fast spin echo, TR/TE = 12300/17.3 ms, interpolated voxel size = 0.49 x 0.49 x 0.8 mm3.

Repetition time (TR), echo time (TE), inversion time (TI), T1 weighted (T1w), T2 weighted (T2w), Fluid-attenuated inversion recovery (FLAIR).

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.

Cohort 1, SABRE: University College London
Cohort 2, RSS: Erasmus Medical Center

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).

Trained radiographers acquired the data according to a predefined research protocol.

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.

Both training and test cases represent coregistered T1w, T2w and FLAIR MRI scans (registered to the T1w MRI scan) of a human brain. Both training and test cases contain two sets of full annotations (as in segmentations) of lacunes in the whole brain (two raters).

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

Train: 40 cases (6 cases from SABRE, 34 cases from RSS), representing in total 40 subjects.
Test: 66 cases (10 cases from SABRE, 56 cases from RSS, same ratio as in training set), representing in total 66 subjects.

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

The number of cases is limited by the prevalence of lacunes in the population in addition to the effort of the labeling task and the number of cases we could make publicly available. The ratio between training and test set was applied similarly to all centers providing data and raters providing annotations (approx. 40% train, 60% test).

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 training and test cases were chosen to represent the variability of number of lacunes and subject age. We applied stratified random sampling aimed at acquiring an approximately uniform distribution over the number of lacunes and the age of the subjects. The training set is a representative set of the test set.

The training set subjects of SABRE overlap with 6 of the cases in the microbleed task and all of the subjects in the enlarged PVS task. The RSS subjects of the training set are the same for all tasks.

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.

All cases contain two sets of segmentations (from two raters). SABRE cases and RSS cases were segmented by two different raters.

Train:
6 scans from SABRE with segmentations in the full brain.
34 scans from RSS with segmentations in the full brain.

Test:
10 scans from SABRE with segmentations in the full brain.
56 scans from RSS with segmentations in the full brain.

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

For both the SABRE and RSS cases the STRIVE criteria (see reference section) were used:
- "A round or ovoid, subcortical, fluid-filled cavity (signal similar to CSF) of between 3 mm and about 15 mm in diameter, consistent with a previous acute small subcortical infarct or haemorrhage in the territory of one perforating arteriole." [6]
- "On fluid-attenuated inversion recovery (FLAIR) images, lacunes of presumed vascular origin generally have a central CSF-like hypointensity with a surrounding rim of hyperintensity; however, the rim is not always present" [6]

Emphasis was given on providing 3D consistent segmentation and on looking at the three modalities simultaneously. Only the lacune, so the fluid-filled cavity, was included in the segmentations. Any surrounding gliosis (the hyperintense rim on FLAIR) was not included in the segmentations. Only lacunes in the cerebrum were

included; lacunes in the cerebellum were excluded as the underlying pathology of lacunes is presumably different in this part of the brain.

One set of annotations for the RSS cases (done by rater 6) was annotated using contours that were subsequently converted to voxelwise labels. The rater generally annotated the lacunes on T1, however for some of the cases lacunes were annotated on the FLAIR scan. The other set (done by rater 7) was annotated voxelwise on the T1 scan.

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.

SABRE: One medically trained rater (rater 1) and one rater with 5+ years of professional experience segmented the lacunes (rater 2). Software based correction was applied on the segmentation to ensure relevance of the signal intensity.

RSS: Several expert raters indicated for all cases if a lacune was present or not (every case was looked at by one expert). Subsequently, for all cases where a lacune was present, one expert rater segmented all lacunes in all cases (rater 6). One medical doctor (4 years of experience with SVD markers) also segmented lacunes in all cases without knowing which cases contained a lacune or not according to other raters (rater 7).

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.

Annotations of two raters will be made available for all cases.

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.

Rigid coregistration of T1w, T2w and FLAIR MRI scans to the T1w MRI scan.

All RSS MRI scans were nonuniformity corrected (NU) with the MINC N3 package

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.

Possible sources of error with respect to annotating lacunes and enlarged PVS in general are the difficulty of distinguishing these markers from each other and similarly appearing structures as well as the lack of a way to uncover the "real" ground truth.
Inter-intra variability is present but the core of this task is to be able to acknowledge and recognize certain and uncertain examples. Inter-rater variability will be assessed on the test set.

A possible source of error is the difficulty of identifying lacunes and differentiating it from mimics. Another source lies in identifying the border of the element, due to partial volume effects it can be difficult to see which voxels are part of the lacune and which are not. This is the case in axial slices, but also in depth, to see until which slice the

lacune is still visible.
Further inacurracies may be due to issues in the use of the segmentation software tool (too large brush, not considering all orientations...) as well as initial misalignment. For the RSS cases (the annotations done by rater 6) a possible error source is the conversion from contours (done by the rater) to voxelwise annotations (segmentation mask).

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

Another source of error may come from the rigid registration of the scans to the T1w MRI scan.

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)

We will compute the following metrics for cases with manual segmentations (averaged over non-empty cases):

Segmentation:Average Dice Similarity Coefficient (DSC) over detected elements
Volume: Absolute Difference
Detection: Detection F1
Uncertainty: DSC for Segmentation Uncertainty, Detection F1 for Detection Uncertainty

Defined as:
DSC (voxelwise), Detection F1 (elementwise) = 2TP / (2TP + FN + FP)
Absolute Difference = $|$Total Predicted - Total Rater$|$ = $|$FP - FN$|$
Uncertainty evaluation terms:
TP_uncertainty: incorrect and uncertain
TN_uncertainty: correct and certain
FP_uncertainty: correct and uncertain
FN_uncertainty: incorrect and certain

We will compute the following metrics for cases with no manual segmentations, so for all empty cases (averaged over empty cases):

Volume: Absolute difference
Uncertainty: DSC for Segmentation Uncertainty, Detection F1 for Detection Uncertainty

Defined as:
Absolute Difference = $|$Total Predicted - Total Rater$|$ = $|$Total Predicted$|$ = $|$FP$|$
Uncertainty metrics are evaluated as defined above for the non-empty cases.

All metrics will be computed based on 1 prediction threshold, namely at 0.5. Detections will be evaluated using a distance threshold (distance between centers of mass).

All test cases have segmentations by 2 raters. The volume metric will be averaged over the 2 raters. Segmentation and detection metrics will be weighed with the raters agreement:
- Agreement: voxel/element has weight = 1
- Disagreement: voxel/element has weight = 0.5

All metrics will be used to compute the ranking for this task. All metrics except the uncertainty metrics will be taken into account for the overall ranking over all 3 tasks.

TP: true positives, FP: false positives, FN: false negatives

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

Submitted methods are expected to output a soft segmentation and an uncertainty map. We apply a threshold of 0.5 on the predicted soft segmentations. We compute the defined metrics for all test cases and then average over these cases.

In the evaluation of this task we focus on four important aspects of lacune annotations, namely segmentation, volume, detection and uncertainty.

Segmentation:
There is an extreme imbalance between background and foreground voxels for lacune segmentation. For this reason we compute DSC to evaluate segmentation performance, this is a commonly used metric that handles imbalance well. DSC is defined as two times the intersection between the prediction and manual segmentations (the number of voxels that are correctly predicted as lacune, true positives) divided by the total number of positive voxels (lacune voxels in manual segmentations) and the total number of voxels predicted to be positive (lacune voxels in prediction). It varies between 0 (no lacune voxels correct) and 1 (all lacune voxels are correctly predicted and no missed or wrong voxels).

Detection:
We will evaluate the detection performance using the commonly used Detection F1. This metric is equivalent to DSC, but is computed at the element level while DSC is computed at the voxel level. Elements are true positive when the distance between the centers of mass is below a chosen threshold (to be announced). This metric varies between 0 (no lacunes detected correctly) and 1 (all lacunes are correctly detected and no missed or wrong detections).

Volume:
We will evaluate the total predicted volume using the absolute difference between the predicted total volume and the rater's total segmented volume. The total (predicted) volume per case is the sum over all foreground voxels in the (predicted) segmentation mask.

Uncertainty
Predicted uncertainty maps are expected to range between 0 (certain) and 1 (uncertain) and should have the same shape as the predicted segmentations masks.
We define the following uncertainty evaluation terms as proposed by Mobiny et al. [10]:
TP_uncertainty: incorrect and uncertain

TN_uncertainty: correct and certain
FP_uncertainty: correct and uncertain
FN_uncertainty: incorrect and certain
Incorrect is FPs and FNs, correct is TPs, certain is defined as < threshold, uncertain is defined as >= uncertainty threshold (to be announced).
We compute the Detection F1 score using these terms to quantify the detection uncertainty. We define the uncertainty value per element as the minimum value (=highest certainty) in the uncertainty map within that element (its predicted segmentation).
The segmentation uncertainty will be quantified using DSC also using these terms, per voxel instead of per element. The segmentation uncertainty will only be computed over TP detections (elements).

Elements (relevant for detection and detection uncertainty metric):
Elements are extracted from the segmentation masks with connected component labeling. We use a connectivity of 6 (only including direct neighbors).

Multiple raters (relevant for segmentation, detection and volume metrics):
For all test cases we have segmentations of two different raters.
- The volume metric will be computed per rater and averaged over the raters.
- The segmentation (DSC) and detection (Detection F1) metrics are weighed by the raters agreement. For segmentation: voxels that the raters agree on have a weight of 1, voxels that raters disagree on have a weight of 0.5. The weighed metrics quantify the similarity between the predictions and the segmentations weighed by how certain the raters are per voxel. The same weighing is applied for the detection metric, however per element instead of per voxel.

For empty cases (no lacunes in scan):
When cases have no lacunes (empty cases) there can be no TPs or FNs, so DSC and Detection F1 are not computed for these cases. The rater's volume is 0 for these cases, so the absolute difference is equal to the total predicted volume, which in this case is the number of false positive voxels.

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.

In order to achieve to final ranking, we will apply the ranking methods described in the Medical Decathlon challenge. For each given metric, and for each participant, the score attributed is the number of other participants performing significantly worse than the given participant. The rank of each participating method is calculated over this score (the higher, the better). The sum of these ranks across metrics is then used as the overall performance, with the lowest rank corresponding to the best performing algorithm.
All sections of the evaluation count equally in the overall ranking, so all metrics in a section have a weight of 1 divided by the number of metrics in that section. We will weigh metrics for non-empty cases and empty cases with the corresponding number of cases.

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

For submissions that result in missing outputs, all missing cases will be given the worst corresponding metrics value if the associated metrics is bounded and a 10% worse value than the worst observed metric across all other

participants for the corresponding case if the metric is not bounded.

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

This is to date the most robust way of providing unbiased ranks when deciding on a collection of metrics.

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.

Wilcoxon signed-rank test.
Bootstrapping test set, computed confidence interval.

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

We will use bootstrapping to simulate the performance of methods for a new set drawn from the distribution. We will use the Wilcoxon signed-rank test to test significance as we will have paired data and the error distributions will probably not be normally distributed. The Wilcoxon signed-rank test is a nonparametric statistical test for paired data.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Further analyses will include inter algorithm variability in terms of the methods and clustering of results. Performance of ensembling methods based either on all or the top 50% will be assessed. Methods will be assessed by performance per center and vascular burden to assess any existing bias.

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.

Cerebral Small Vessel Disease markers:
[1] Cuadrado-Godia, Elisa, et al. "Cerebral small vessel disease: a review focusing on pathophysiology, biomarkers, and machine learning strategies." Journal of stroke 20.3 (2018): 302.

Southall and Brent Revisited (SABRE) dataset - overall study and wave3 data:
[2] Tillin T, Forouhi NG, McKeigue PM, Chaturverdi N for the S group, Chaturvedi N. Southall And Brent REvisited: cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. Int J Epidemiol 2012; 41: 33–42.

[3] Sudre CH, Smith L, Atkinson D, Chaturvedi N, Ourselin S, Barkhof F, et al. Cardiovascular Risk Factors and White Matter Hyperintensities: Difference in Susceptibility in South Asians Compared With Europeans. J Am Heart Assoc 2018; 7

Rotterdam Scan Study (RSS) dataset, large population study:
[4] Ikram, M. Arfan, et al. "Objectives, design and main findings until 2020 from the Rotterdam Study." European Journal of Epidemiology (2020): 1-35.

ALFA dataset
[5] Salvadó G, Brugulat-Serrat A, Sudre CH, Grau-Rivera O, Suárez-Calvet M, Falcon C, et al. Spatial patterns of white matter hyperintensities associated with Alzheimer's disease risk factors in a cognitively healthy middle-aged cohort. Alzheimers Res Ther 2019; 11: 12.

STRIVE criteria, annotating SVD biomarkers (lacunes, pvs, microbleeds):
[6] Wardlaw, Joanna M., et al. "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration." The Lancet Neurology 12.8 (2013): 822-838.

UNIVRSE rating system, annotating enlarged PVS (RSS cases):
[7] Adams, Hieab HH, et al. "Rating method for dilated Virchow–Robin spaces on magnetic resonance imaging." Stroke 44.6 (2013): 1732-1735.

BOMBS criteria, annotating microbleeds:
[8] Cordonnier, C., Potter, G.M., Jackson, C.A., Doubal, F., Keir, S., Sudlow, C.L.M., Wardlaw, J.M., Al-Shahi Salman, R., 2009. Improving interrater agreement about brain microbleeds: Development of the Brain Observer MicroBleed Scale (BOMBS). Stroke 40, 94–99.

Annotating microbleeds (RSS cases):
[9] Vernooij, M. W., et al. "Prevalence and risk factors of cerebral microbleeds: the Rotterdam Scan Study." Neurology 70.14 (2008): 1208-1214.

Uncertainty metrics:
[10] Mobiny, Aryan, et al. "DropConnect Is Effective in Modeling Uncertainty of Bayesian Deep Networks." arXiv preprint arXiv:1906.04569 (2019). (see section III E of this paper)
[11] Camarasa, Robin, et al. "Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation." UNSURE 2020, Springer (2020): 32-41.

Ranking scheme Medical Decathlon:
[12] http://medicaldecathlon.com/files/MSD-Ranking-scheme.pdf

Further comments

Further comments from the organizers.

With this challenge we would like to encourage more people to work on this application and increase awareness of the interesting challenges in this application. We would like to host this challenge in 2021, so we will have enough time to promote this challenge and provide people that are new to this application enough time to be able to participate.