VAscular Lesions DetectiOn: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

VAscular Lesions DetectiOn

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 are one such prominent marker of cerebral small vessel disease 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 sparsity in the data but also the absence of sufficient gold standard/and or evaluation of the labeling noise.

This challenge aims at promoting the development of new solutions for the automated detection, differentiation and segmentation of such very sparse and small objects while leveraging the presence of weak and noisy labels.

The challenge will have a technical impact in the following fields: use of weak labels, assessment of prediction uncertainty, object detection, class imbalance, multi-scale object detection. 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 (mammography, lung nodule detection...).

Challenge keywords

List the primary keywords that characterize the challenge.

Extremely small objects – detection – segmentation - noisy labels – vascular – perivascular spaces - microbleeds - lacunes - co-occurrence

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 are hoping for 15 participants. We expect at least 5 specific groups to participate (MedDigit, University of Magdeburg, confirmed; Image Sciences Institute, Utrecht University, possible; Center for Neurological Imaging, Brigham and Women's Hospital, possible; Biomedical Research Imaging Center, University of North Carolina, possible; Laboratory of NeuroImaging, University of Southern California, possible). In addition we will send an inquiry using medical imaging community mailing lists to further evaluate the interest in the community. 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 Grand Challenge platform will be used for the submission of solutions. Evaluation of the submissions will be done partly at the Erasmus Medical Center and King's College London as part of the test set data cannot leave these centers and if possible partly on the GPUs supplied by NVIDIA.

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 this 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.
Furthermore, manual annotation is subject to observer bias due to the difficulty of distinguishing an enlarged PVS from other similarly appearing structures.
Dealing with this subjectivity of annotations is the main challenge for current automated methods, as it is not possible to acquire a "real" ground truth.
A robust, automated method for segmenting PVS would be extremely useful for neurological research on the role of enlarged PVS in neurological disorders. It would further allow the identification of relevant quantifiable measures that could be subsequently derived from the segmentation.

Keywords

List the primary keywords that characterize the task.

Enlarged PVS - detection - segmentation - noisy labels

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 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

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

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

Open call.

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.

We intend to use Grand-challenge.org.

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

TBA

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, even those available for other tasks. Using private datasets for training is discouraged and these submissions will not be eligible for awards. Participants have to indicate whether they used additional private training data. This information will be displayed on the leaderboard.
Participants who want to use their own training data and still be eligible for awards will need to share this data with the other participants. We will mediate this process by performing a quality check of the data and make it available for download on the challenge platform.

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. There will be an award for each winner of each task and 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. 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 website.

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, the first and last 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.

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 subset of the test set before their final submission. This will be allowed twice at most. Their final submission will be evaluated on the full 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

November 2020: Release of training data
End of March 2021: Optional deadline for participants to send their additional training data for use in the challenge (see section training data policy, part b)
Mid April 2021 - Mid July 2021: Optional submission for evaluation on a small subset of the test set (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: Link 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 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 and/or annotations.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, Screening, Diagnosis, Prognosis.

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

Imaging modality(ies)

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

MRI - T1, T2, FLAIR

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 - T1, T2, 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)).

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

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): Philips 3T
Cohort 2, Rotterdam Scan Study (RSS): 1.5T MRI scanner General Electric Healthcare (GE)

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).

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 scan) of a human brain. Training cases present full annotation of a subset of axial slices in three discontinuous locations. Test cases either correspond to a fully annotated brain or a subset of axial slices.

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

Train: 26 subjects with annotated slabs and slices, containing around 2000 elements annotated in total (6 SABRE with rater 1 and rater 2, 20 RSS with rater 3).
Test: 50 subjects (10 SABRE with rater 1 and rater 2, 23 RSS with rater 3, 12 RSS with rater 3 and rater 4, 5 RSS with rater 5).

At this time this is the maximum amount we can commit to but we are currently looking into ways of further expanding this dataset without compromising its very high quality.

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

The number of training cases was decided based on the limited number of annotated scans with segmentations.
The ratio defined between training and test set was applied similarly to all centers providing data and raters providing annotations.
The test set contains 10 SABRE scans with annotations by rater 1 and 2 and 35 RSS scans with annotations by rater 3, this is the same ratio as in the training set. More RSS scans are included in the test set with annotations by rater 4 and rater 5.
Annotations of multiple raters are used in the test set to evaluate how well methods handle different and unseen rater styles.
We estimate that there are around 20 segmentations per slice and about 50 segmentations per slab. There will be 3 slabs with segmentations provided per scan for SABRE and 3 slices with segmentations provided per scan for RSS. This adds up to roughly 2000 elements in the training set.

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 scans are chosen so the range of numbers of enlarged PVS that can be encountered are represented. Scans will range from having no enlarged PVS, to many enlarged PVS. The test set is chosen in the same way.
The 26 subjects in the training set are also part of the training set in task 2 and 4. The test set subjects do not overlap with any of the training set subjects of the 4 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 scans from SABRE with manual segmentations of rater 1 and rater 2 in slabs in order to get tridimensional information.
20 scans from RSS with manual segmentations from rater 3 on 3 slices.
In total the training set contains around 2000 segmentations.

Test:
10 scans from SABRE with manual segmentations of rater 1 and rater 2 in the full brain.
12 scans from RSS with manual segmentations of rater 3 and rater 4 in slices.
23 scans from RSS with manual segmentations of rater 3 in slices.
5 scans from RSS with manual segmentations of rater 5 in one hemisphere.

At this time this is the maximum amount we can commit to but we are currently looking into ways of further expanding this dataset without compromising its very high quality.

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.

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

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: 1 Medically trained rater and 1 with 5+ years of professional experience - software based correction was applied on the segmentation to ensure relevance of the signal intensity (raters 1 and 2).
RSS: segmentations done by medical students trained to recognize enlarged perivascular spaces (rater 3 and 4) and an expert rater (rater 5).

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 multiple 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.

Affine coregistration of T1, T2 and FLAIR images to T1 space. We will provide code to facilitate the processing of slabs of brain scans surrounding the annotated slices. Slices are distributed throughout the brain so that the whole coverage of the brain morphology will be presented in training slabs.

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 distinguishing enlarged PVS from similarly appearing structures as well as the lack of a way to uncover the "real" ground truth. Furthermore 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.
As of date - the second rater is still creating the segmentation on the SABRE data - inter rater variability will only be evaluated afterwards.
Inter-rater variability will be assessed on a subset of the test cases.
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

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

Another source of error may come from the registration of the scans to the T1 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)

Probabilistic Dice Similarity Coefficient (detection + volumetry)
Distance to center of mass (detection)
Detection F1 (detection)
Sensitivity (element level) (detection)
Specificity (element level) (detection)
FROC (detection)
Elementwise volume correlation (volumetry + detection)
Volume difference ratio (volumetry)

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

At the level of enlarged PVS detection, both detection and segmentation are important aspects that must be covered by the evaluation metrics. In terms of local detection, measure of overlap will be used, which will be re-weighted based on multi-rater agreement (Probabilistic dice overlap). Due to the very limited size of the objects to be found, the distance to the corresponding object center of mass will be used. Assessment of overall volume with the ratio of the difference with respect to reference segmentation will be included. At the detection level, F1 score, sensitivity and specificity at the level of the connected component will be used (detection assessment). In addition an elementwise volume correlation will be included to evaluate simultaneously volumetry and detection.
The inter rater agreement will be calculated and disagreements will be down weighed. Each voxel for which disagreement is recorded will be given a weight of 0.5 (instead of 1) in the integration of the maps of true positives, false negatives and false positives. A similar approach will be used at the component level if disagreement exist at that level. The methods will be evaluated using raters' annotations separately and the performance on all annotation sets will be combined. The single rater evaluations will be compared to the inter-rater variability measurement.

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.

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 corresponding case.

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. Evaluation 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: Counting of enlarged PVS in axial slices

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.

Currently, neurological studies mostly score the burden of enlarged PVS visually by e.g. counting the number of enlarged PVS in a slice, as this is the most practical and fast way to quantify enlarged PVS in a scan. For large datasets this still very time-consuming and can be subjective due to the difficulty of distinguishing an enlarged PVS from other similarly appearing structures. An automated method for quantification of PVS burden has the potential to be faster and more objective than manual counting, making it possible to obtain larger annotated datasets for neurological research.

Keywords

List the primary keywords that characterize the task.

Enlarged PVS - weak label - count - visual score

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 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.

Examples:

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

Open call.

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.

We intend to use Grand-challenge.org.

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

TBA

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, even those available for other tasks. Using private datasets for training is discouraged and these submissions will not be eligible for awards. Participants have to indicate whether they used additional private training data. This information will be displayed on the leaderboard.
Participants who want to use their own training data and still be eligible for awards will need to share this data with the other participants. We will mediate this process by performing a quality check of the data and make it available for download on the challenge platform.

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. There will be an award for each winner of each task and 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. 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 website.

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, the first and last 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.

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 subset of the test set before their final submission. This will be allowed twice at most. Their final submission will be evaluated on the full 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

November 2020: Release of training data
End of March 2021: Optional deadline for participants to send their additional training data for use in the challenge (see section training data policy, part b)
Mid April 2021 - Mid July 2021: Optional submission for evaluation on a small subset of the test set (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: Link 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 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 and/or annotations.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, Screening, Diagnosis, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Regression.

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 is a tri-ethnic ageing population with high cardiovascular risk factors.
Cohort 2 is a population study of an ageing population in a homogeneous environment.
We have the intention of adding another cohort.

Imaging modality(ies)

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

MRI - T1, T2, FLAIR

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 - T1, T2, 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)).

Reliability, Specificity, Sensitivity, Precision, Accuracy.

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): Philips 3T
Cohort 2, Rotterdam Scan Study (RSS): 1.5T MRI scanner General Electric Healthcare (GE)

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).

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 scan) of a human brain. Both training and test cases have counts for selected slices.

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

Train: 40 subjects (6 SABRE, 34 RSS).
We have the intention to add another set with about 12 scans with visual scores to the training set.
Test: 66 subjects (10 SABRE, 56 RSS, same ratio as in training set).
We have the intention to add another set with about 20 scans with visual scores to the test set (would be same set as the training set).

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

Manually annotating scans by counting enlarged PVS is far less time-consuming than segmenting PVS, so more training cases are available in task 2 than in task 1. Training and test proportions were adopted so as to maintain similar ratio balance in training and test set across centers providing data.

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 scans are chosen so the range of numbers of enlarged PVS that can be encountered are represented. Scans will range from having no enlarged PVS, to many enlarged PVS. The test set is chosen in the same way.
The 40 subjects in the training set are the same 40 subjects as in the training set in task 4, and 26 of these are also part of the training set in task 1. Per task annotations for different slices will be provided however. We have the intention to add another dataset to the train and test set with visual scores which will have no subjects that overlap with other tasks. The test set subjects do not overlap with any of the training set subjects of the 4 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.

RSS: the number of PVS in a slice manually counted by medical doctor following published rating protocol (UNIVRSE).
SABRE: counts computed from scans annotated with manual segmentations.

Train:
6 scans from SABRE with counts in slices (rater 1 and rater 2).
34 scans from RSS with counts in slices (rater 3).
We have the intention to add another set with about 12 scans with counts.

Test:
10 scans from SABRE with counts in slices (rater 1 and rater 2).
66 scans from RSS with counts in slices (rater 3 and we aim to have a rater 4).
We have the intention to add another set with about 20 scans with counts.

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.

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.

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: 1 Medically trained rater and 1 with 5+ years of professional experience (raters 1 and 2).
RSS: medical doctor following published rating protocol (UNIVRSE) (rater 3) and we aim to add a 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.

Annotations of multiple 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.

Affine coregistration of T1, T2 and FLAIR images to T1 space. We will provide code to facilitate the processing of slabs of brain scans surrounding the annotated slices. Slices are distributed throughout the brain so that the whole coverage of the brain morphology will be presented in training slices.

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 distinguishing enlarged PVS from similarly appearing structures as well as the lack of a way to uncover the "real" ground truth. Furthermore enlarged PVS can be very small and easy to miss.
Possible sources of error with respect to visual scores relate to the choice of the slice on which to perform the assessment as well as possible observational shift. However the scale used has been shown to present a strong inter-rater consistency. Furthermore, the existence of mimics can be a source of error. Intra-rater variability will be assessed on a subset of the test cases.

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

Another source of error may come from the registration of the scans to the T1 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)

Root mean square error between rater's count and predicted count
Kendall's tau correlation between rater's count and predicted count
Intra-class correlation (ICC) between rater's count and predicted count

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

The metrics quantify how well the count is predicted. ICC is the most used metric in literature when assessing PVS counts and visual scoring. The predictions of the methods will be compared to the counts of the raters and the counting difference will be weighed by the number of lesions that are present in the image. Counting a few more lesions is worse when only a few lesions are present in the scan than when there are many.

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.

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 corresponding case.

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. Evaluation 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. Currently microbleeds are identified manually and only their count is considered in association studies. Incorporating size information could be relevant, but this would require an appropriate delineation. Microbleeds presence has been associated with the existence of specific vascular pathology such as cerebral amyloid angiopathy and other markers of cerebral small vessel disease (WMH, enlarged PVS). The challenges of automated identification of microbleeds are the presence of numerous mimics and the sparsity of the data both in terms of the size of the element of interest and their distribution (rarely more than 5 in a scan). An automated method for microbleed segmentation would notably enable further research on their presence in the context of neurodegenerative diseases.

Keywords

List the primary keywords that characterize the task.

Cerebral microbleeds - detection - segmentation

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 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

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

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

Open call.

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.

We intend to use Grand-challenge.org.

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

TBA

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, even those available for other tasks. Using private datasets for training is discouraged and these submissions will not be eligible for awards. Participants have to indicate whether they used additional private training data. This information will be displayed on the leaderboard.
Participants who want to use their own training data and still be eligible for awards will need to share this data with the other participants. We will mediate this process by performing a quality check of the data and make it available for download on the challenge platform.

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. There will be an award for each winner of each task and 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. 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 website.

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, the first and last 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.

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 subset of the test set before their final submission. This will be allowed twice at most. Their final submission will be evaluated on the full 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

November 2020: Release of training data
End of March 2021: Optional deadline for participants to send their additional training data for use in the challenge (see section training data policy, part b)
Mid April 2021 - Mid July 2021: Optional submission for evaluation on a small subset of the test set (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: Link 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 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 and/or annotations.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, Screening, Diagnosis, Prognosis.

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 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).
Cohort 2 is a tri-ethnic ageing population with high cardiovascular risk factors.
Cohort 3 is a population study of an ageing population in a homogeneous environment.

Imaging modality(ies)

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

MRI - T1, SWI or T2*

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 - T1, SWI or T2*

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)).

Reliability, Specificity, Sensitivity, Precision, Accuracy.

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, ALFA: GE Discovery 3T
Cohort 2, Southall and Brent Revisited (SABRE): Philips 3T

Cohort 3, Rotterdam Scan Study (RSS): 1.5T MRI scanner General Electric Healthcare (GE)

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

Cohort 1, 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.

Cohort 2, 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.

Cohort 3, 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 = 0.5 x 0.5 x 0.8 mm3.

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

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, ALFA: Barcelona Brain Research Center
Cohort 2, SABRE: University College London
Cohort 3, 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 and SWI or T2*w acquisition (registered to the T1w scan) of a human brain. Both training and test cases have the associated full annotation of cerebral microbleeds.

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

Train: 74 subjects (10 SABRE, 34 RSS, 30 scans of the ALFA study).
Test: 148 subjects (20 SABRE, 68 RSS, 60 scans of the ALFA study).

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 microbleeds in the population in addition to the difficulty of the labeling task. Training and test proportions were adopted so as to maintain similar ratio balance in training and test set across centers providing data.

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 scans are chosen so the range of numbers of microbleeds that can be encountered are represented. The test set is chosen in the same way.
The 34 RSS subjects in the training set are the same 34 subjects as in the training set in task 4, and 26 of these are also part of the training set in task 1. In this task we provide T2* MRI scans with annotations that are registered to the T1 scan (same as for the other tasks). The subjects of the ALFA study do not overlap with the other tasks. 6 of the SABRE subjects in the training set overlap with task 1, 2 and 4, the other 4 SABRE subjects in the training set do not overlap with the training nor test set of the other tasks. The test set subjects do not overlap with any of the training set subjects of the 4 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.

ALFA and SABRE: Manual segmentation by two human raters supervised by an expert neuroradiologist.
RSS: manual segmentation by an expert.

Train:
30 scans from ALFA with segmentations.
10 scans from SABRE with segmentations.
34 scans from RSS with segmentations.

Test:
60 scans from ALFA with segmentations.
20 scans from SABRE with segmentations.
68 scans from RSS with segmentations.

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

BOMBS criteria (see reference section).

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.

ALFA: 1 Medically trained rater (rater 1 from task 1 and 2) and 1 rater in training performed the segmentation and were supervised by a neuroradiologist with 10 + years of professional experience.
RSS: segmentations were performed by an expert rater.

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.

Affine coregistration of T1 and SWI images to SWI space.

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 source of errors in the annotation is the existence of mimics to cerebral microbleeds as well as the definition of the elements borders.

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

Another source of error may come from the registration of the scans to the T1 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)

Dice Similarity Coefficient (detection + volumetry)
Distance to center of mass (detection)
Detection F1 (detection)
Sensitivity (element level) (detection)
Specificity (element level) (detection)
FROC (detection)
Elementwise volume correlation (volumetry + detection)
Volume difference ratio (volumetry)

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

Similarities in the lesion size properties leads to the adoption for consistency purposes of the same metrics as those used for the Task 1.
The inter rater agreement will be calculated and voxels of disagreements will be down weighed in the assessment.

Each voxel for which disagreement is recorded will be given a weight of 0.5 (instead of 1) in the integration of the maps of true positives, false negatives and false positives. A similar approach will be used at the component level if disagreement exist at that level. The methods will be evaluated using raters' annotations separately and the performance on all annotation sets will be combined. The single rater evaluations will be compared to the inter-rater variability measurement.

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.

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 corresponding case.

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. Evaluation 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: Probabilistic classification of enlarged PVS and 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.

Enlarged PVS and lacunes are both markers of cerebral small vessel disease. They have a similar intensity and can have a similar shape in MRI scans. They are often mistaken for each other, even for experts distinguishing enlarged PVS from lacunes can be challenging and sometimes impossible depending on the MRI scan quality. The inter-rater variability is high for the classification between enlarged PVS and lacunes. This variability should be modeled when automatically classifying elements across these two classes. This would avoid errors in quantification of small vascular markers due to double counting, e.g. considering the same element as both lacune and enlarged PVS). The challenge will contribute to the development of techniques modeling labeling uncertainty and label noise in the context of cerebral small vessel disease markers.

Keywords

List the primary keywords that characterize the task.

Probabilistic classification - noisy labels - uncertainty - lacunes vs enlarged PVS

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 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

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

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

Open call.

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.

We intend to use Grand-challenge.org.

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

TBA

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, even those available for other tasks. Using private datasets for training is discouraged and these submissions will not be eligible for awards. Participants have to indicate whether they used additional private training data. This information will be displayed on the leaderboard.
Participants who want to use their own training data and still be eligible for awards will need to share this data with the other participants. We will mediate this process by performing a quality check of the data and make it available for download on the challenge platform.

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. There will be an award for each winner of each task and 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. 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 website.

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, the first and last 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. The organizing team will apply these to the test set to evaluate the performance on the test set.

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 subset of the test set before their final submission. This will be allowed twice at most. Their final submission will be evaluated on the full 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

November 2020: Release of training data
End of March 2021: Optional deadline for participants to send their additional training data for use in the challenge (see section training data policy, part b)
Mid April 2021 - Mid July 2021: Optional submission for evaluation on a small subset of the test set (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: Link 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 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 and/or annotations.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, Screening, Diagnosis, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

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

Imaging modality(ies)

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

MRI - T1, T2, FLAIR

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 - T1, T2, 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) and lacunes.

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)).

Reliability, Specificity, Sensitivity, Precision, Accuracy.

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): Philips 3T

Cohort 2, Rotterdam Scan Study (RSS): 1.5T MRI scanner General Electric Healthcare (GE)

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).

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 scan) of a human brain. For both training and test sets, a csv file with the voxel location of the center of mass of the object of interest, its extent and its associated probabilistic classification is provided.

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

Train: 40 subjects with around 3000 elements annotated in total (6 SABRE, 34 RSS).
Test: 66 subjects (10 SABRE, 56 RSS, same ratio as in training set).

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

Annotating the location of enlarged PVS is far less time-consuming than segmenting PVS, so more training cases are available in task 4 than in task 1. Training and test proportions were adopted so as to maintain similar ratio balance in training and test set across centers providing data.
We estimate that there are around 20 segmentations per slice and about 50 segmentations per slab of 3 slices. There will be 3 slabs with segmentations provided per scan for SABRE (6 scans in total) and 3 slices with segmentations provided per scan for RSS (34 scans in total). This adds up to roughly 3000 elements in the training set, 50 of which expected to be lacunes. Lacunes are effectively seen much more rarely than enlarged perivascular spaces but their distinction is sometimes difficult.

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 scans are chosen so the range of numbers of enlarged PVS that can be encountered are represented. This is also the case for lacunes. The test set is chosen in the same way.
The 40 subjects in the training set are the same 40 subjects as in the training set in task 4, and 26 of these are also part of the training set in task 1. Per task annotations for different slices will be provided however. The test set subjects do not overlap with any of the training set subjects of the 4 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.

6 raters classify objects of interest based on a zoomed in view in a GUI setting as either an enlarged perivascular space, a lacune or as neither of these. For all scans a csv file with the voxel location of the center of mass of the object of interest, its extent and its associated probabilistic classification is provided.

Train:
6 scans from SABRE with objects of interest in 3 slabs.
34 scans from RSS with objects of interest in 3 slices.
In total the training set contains around 3000 objects of interest.

Test:
10 scans from SABRE with objects of interest in slices.
56 scans from RSS with objects of interest in slices.

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

The 6 participating raters underwent training in the definition of the objects of interest and in the use of software GUI tool used to store their ratings. For each presented element, raters had to decide if the element was an enlarged perivascular space, a lacune or nothing.

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 and RSS: 1 neuroradiologist with 10+ years, 4 medically trained rater with 3-5 years experience and 1 biomedical engineer with 5+year of experience performed the classification.

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.

Per class probability will be given as the average of the different raters using a majority voting classification.

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.

Affine coregistration of T1, T2 and FLAIR images to T1 space. We will provide code to facilitate the processing of slabs of brain scans surrounding the annotated slices. Slices are distributed throughout the brain so that the whole coverage of the brain morphology will be presented in training slabs.

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.

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

Another source of error may come from the registration of the scans to the T1 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)

RMSE on classification probability
Intra-class correlation
Confusion matrix metrics on classification performance
Confusion matrix metrics on classification uncertainty

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

The task focuses on classification and uncertainty assessment. The metrics have been chosen to assess how well the overall class was chosen as well as reproduce the level of uncertainty of the rating. The objective is to predict this probability as surrogate of classification uncertainty, which is why the RMSE over the probabilities is taken as evaluation measure.
Confusion matrix metrics on uncertainty will focus on evaluating purely uncertainty without taking into account classification performance. These definitions will be used:
TP: incorrect and uncertain
TN: correct and certain
FP: correct and uncertain
FN: incorrect and certain
See section 'additional points' for the paper explaining this.

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.

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 corresponding case.

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. Evaluation 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:
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.

Review automated methods to quantify enlarged PVS:
del C. Valdés Hernández, Maria, et al. "Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: a systematic review." Journal of Magnetic Resonance Imaging 38.4 (2013): 774-785.

Southall and Brent Revisited (SABRE) dataset - overall study and wave3 data:
- Tillin T, Forouhi NG, McKeigue PM, Chatuverdi 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.
- 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:
Ikram, M. Arfan, et al. "The Rotterdam Scan Study: design update 2016 and main findings." European journal of epidemiology 30.12 (2015): 1299-1315.

ALFA dataset
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.

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

BOMBS criteria, annotating microbleeds:
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.

UNIVRSE criteria, annotating PVS counts:
Adams, Hieab HH, et al. "Determinants of enlarged Virchow-Robin spaces: The UNIVRSE consortium." Alzheimer's & Dementia: The Journal of the Alzheimer's Association 10.4 (2014): P408.

STRIVE criteria, annotating SVD biomarkers (lacunes, pvs, microbleeds):
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.

Confusion metrics on uncertainty:
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)

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.