Enlarged Perivascular Spaces (EPVS) Segmentation Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Enlarged Perivascular Spaces (EPVS) Segmentation Challenge

Challenge acronym

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

EPVS Challenge

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.

Our proposed challenge focuses on the segmentation of enlarged perivascular spaces (EPVS) in brain MRIs. EPVS serves as a hallmark and a very early anomaly of cerebral vascular diseases. Previous studies[1-3] have shown its linkage to various neurovascular and neurodegenerative conditions, such as small vessel disease and risk of cognitive impairment. Recently, emerging evidence has shown the early appearance of EPVS in middle-aged population studies and its potential impact on health outcomes such as cognition[4-5].

From both clinical and research standpoints, precise quantification of EPVS is crucial for clinical diagnosis and the assessment of progression in neurological disorders. The current quantification of EPVS still relies on coarse-grained visual scoring [6-7], where clinicians provide a severity index based on the stage of EPVS. This approach is prone to inaccuracies, potential for human error, and high inter-rater variability. Although preferable, fine-grained manual annotation is impractical due to its time-intensive nature. One recent work [8] proposed a 3D U-Net for automatic segmentation of EPVS, trained on a single cohort of young individuals. The researchers reported a performance degradation and a substantial increase in false positives when the model was employed on older populations. This indicates that there remains large room for methodology development and validation for automatic EPVS segmentation.

The technical complexity of accurately identifying EPVS stems from their variability in size, shape, location, and sample, they can appear as small, linear structures, or as larger cystic spaces in the human brain. Moreover, differentiating EPVS from other similar-appearing features like lacunes (small, deep cerebral infarcts) or microbleeds is challenging.

To overcome these challenges, in 2021, the VALDO challenge (https://valdo.grand-challenge.org/) [9] provided a publicly available dataset to encourage participants to develop automatic methods to segment EPVS in MRI scans. Due to the nature of the data scarcity, this challenge provided 12 MRI volumes as a training set, in which only 6 of them were annotated with full slices (the remaining 6 were labeled only for critical slices). This scarcity of annotated data presents a bottleneck for the development of fully automated and generalizable segmentation methods. Notwithstanding the rapid progression in deep learning-based approaches, the winning solution of the challenge leveraged a random forest classifier to surpass competing methodologies, underscoring the opportunity for refinement and innovation of data-driven approaches (especially deep-learning-based) in this field.

In our EPVS challenge, we aim to advance machine learning models that are robust and effective across diverse imaging scenarios, including variations in MRI protocols, resolutions, and demographic profiles from the UK, Singapore, and China. The challenge provides a dataset with 100 fully annotated training scans, 50 validation scans, and 70 scans for independent testing, sourced from multiple sites and scanners to reflect real-world clinical and research conditions (with varying cognitive profiles). The goal is to motivate participants to develop EPVS segmentation algorithms that are adaptable and generalizable, capable of performing well across the broad range of MRI data encountered in clinical practice. This approach ensures that the developed algorithms will be applicable and beneficial across different clinical settings and research studies.

In Summary, by offering data from four distinct sites: National University of Singapore (Singapore), National University Hospital (Singapore), Shanghai University of Traditional Chinese Medicine (China), and University of Edinburgh (UK). We provide a foundation for creating models that can significantly enhance the understanding, early diagnosis, and monitoring of cerebral vascular diseases. This challenge is poised to impact the development of more effective treatments and preventive strategies for neurovascular and neurodegenerative disorders, leveraging advancements in machine learning to improve patient outcomes worldwide.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

EPVS, Segmentation, Brain MRI, Cerebral Vascular Disease

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

N/A

Duration

How long does the challenge take?

Half day.

Expected number of participants

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

We expect 20 participating teams.

Publication and future plans

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

Yes, we intend to coordinate a publication of the challenge results. We aim to compile a comprehensive analysis of the methodologies, results, and insights gained from the challenge. This will include a comparative evaluation of the various approaches adopted by the participants and an assessment of their efficacy and innovation. We plan to submit the consolidated findings to a peer-reviewed academic journal or conference. Additionally, we will encourage collaboration and co-authorship among participants to ensure that their contributions are duly recognized in the 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.

For the online part of our challenge, we're using a custom-built platform. Participants will upload their Docker containers to our website, and we'll handle the evaluations on our end.
Regarding the on-site challenge, we'll need the following equipment to ensure everything runs smoothly on the day of the challenge: projectors or screens for presentations, loudspeakers for clear audio across the venue, and microphones for presentations and discussions.

TASK 1: Enlarged Perivascular Spaces (EPVS) Segmentation Challenge

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.

Our proposed challenge focuses on the segmentation of enlarged perivascular spaces (EPVS) in brain MRIs. EPVS serves as a hallmark and a very early anomaly of cerebral vascular diseases. Previous studies[1-3] have shown its linkage to various neurovascular and neurodegenerative conditions, such as small vessel disease and risk of cognitive impairment. Recently, emerging evidence has shown the early appearance of EPVS in middle-aged population studies and its potential impact on health outcomes such as cognition[4-5].

From both clinical and research standpoints, precise quantification of EPVS is crucial for clinical diagnosis and the assessment of progression in neurological disorders. The current quantification of EPVS still relies on coarse-grained visual scoring [6-7], where clinicians provide a severity index based on the stage of EPVS. This approach is prone to inaccuracies, potential for human error, and high inter-rater variability. Although preferable, fine-grained manual annotation is impractical due to its time-intensive nature. One recent work [8] proposed a 3D U-Net for automatic segmentation of EPVS, trained on a single cohort of young individuals. The researchers reported a performance degradation and a substantial increase in false positives when the model was employed on older populations. This indicates that there remains large room for methodology development and validation for automatic EPVS segmentation.

The technical complexity of accurately identifying EPVS stems from their variability in size, shape, location, and sample, they can appear as small, linear structures, or as larger cystic spaces in the human brain. Moreover, differentiating EPVS from other similar-appearing features like lacunes (small, deep cerebral infarcts) or microbleeds is challenging.

To overcome these challenges, in 2021, the VALDO challenge (https://valdo.grand-challenge.org/) [9] provided a publicly available dataset to encourage participants to develop automatic methods to segment EPVS in MRI scans. Due to the nature of the data scarcity, this challenge provided 12 MRI volumes as a training set, in which only 6 of them were annotated with full slices (the remaining 6 were labeled only for critical slices). This scarcity of annotated data presents a bottleneck for the development of fully automated and generalizable segmentation methods. Notwithstanding the rapid progression in deep learning-based approaches, the winning solution of the challenge leveraged a random forest classifier to surpass competing methodologies, underscoring the opportunity for refinement and innovation of data-driven approaches (especially deep-learning-based) in this field.

In our EPVS challenge, we aim to advance machine learning models that are robust and effective across diverse imaging scenarios, including variations in MRI protocols, resolutions, and demographic profiles from the UK, Singapore, and China. The challenge provides a dataset with 100 fully annotated training scans, 50 validation scans, and 70 scans for independent testing, sourced from multiple sites and scanners to reflect real-world clinical

and research conditions (with varying cognitive profiles). The goal is to motivate participants to develop EPVS segmentation algorithms that are adaptable and generalizable, capable of performing well across the broad range of MRI data encountered in clinical practice. This approach ensures that the developed algorithms will be applicable and beneficial across different clinical settings and research studies.

In Summary, by offering data from four distinct sites: National University of Singapore (Singapore), National University Hospital (Singapore), Shanghai University of Traditional Chinese Medicine (China), and University of Edinburgh (UK). We provide a foundation for creating models that can significantly enhance the understanding, early diagnosis, and monitoring of cerebral vascular diseases. This challenge is poised to impact the development of more effective treatments and preventive strategies for neurovascular and neurodegenerative disorders, leveraging advancements in machine learning to improve patient outcomes worldwide.

Keywords

List the primary keywords that characterize the task.

EPVS, Segmentation, Brain MRI, Cerebral Vascular Disease

ORGANIZATION

Organizers

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

National University of Singapore (NUS) / National University Hospital (NUH), Singapore
Juan Helen Zhou, NUS, Singapore
Christopher Chen, NUS/NUH, Singapore
Yilei Wu, NUS, Singapore
Zijian Dong, NUS, Singapore
Zijiao Chen, NUS, Singapore
Fang Ji, NUS, Singapore
Huijuan Chen, NUS, Singapore
Gifford Tan, NUH, Singapore
An Sen Tan, NUH, Singapore
Sizhao Tang, NUH, Singapore

Harvard Medical School, USA
Hongwei Bran Li, Harvard Medical School, USA

Longhua Hospital Shanghai University of Traditional Chinese Medicine, China
Xin Chen, Longhua Hospital Shanghai University of Traditional Chinese Medicine, China

Technical University of Munich, Germany
Zhenyu Gong, Technical University of Munich, Germany
Benedikt Wiestler, Technical University of Munich, Germany

Centre for Clinical Brain Sciences, University of Edinburgh, UK

Maria del C. Valdés Hernández, Centre for Clinical Brain Sciences, University of Edinburgh, UK
Roberto Duarte Coello, Centre for Clinical Brain Sciences, University of Edinburgh, UK
Joanna M. Wardlaw, Centre for Clinical Brain Sciences, University of Edinburgh, UK
John McFadden, Centre for Clinical Brain Sciences, University of Edinburgh, UK
José Bernal Moyano, Centre for Clinical Brain Sciences, University of Edinburgh, UK; German Centre for Neurodegenerative Diseases (DZNE), Germany

b) Provide information on the primary contact person.

Yilei Wu (yilei.wu@u.nus.edu), Juan Helen Zhou (helen.zhou@nus.edu.sg)

Life cycle type

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

Examples:

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

Open call

Challenge venue and platform

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

MICCAI2024

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

We will build our own challenge website, which will be made publicly available after the acceptance.

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

Will be made publicly available after the acceptance.

Participation policies

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

Fully automatic.

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

The participants can use publicly available data as they wish but should document whatever is used in the description of their algorithm. Participants may modify the provided training data as they wish. This modification includes the generation of additional data by various data augmentation strategies as long as everything is documented, and synthetic data should be able to be made available upon request.

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

May participate but is not eligible for awards.

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

The top three ranking methods will be publicly named and awarded certificates and a small gift.

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 results will be announced publicly at the MICCAI 2024 challenge session, and will be posted on the challenge website. The teams with the top 3 algorithms will be informed earlier so they can prepare a presentation for the challenge session. The top 3 teams will be asked to prepare a 7-10 minute presentation.

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).

Three authors per team who contributed to the design of the algorithm will be named co-authors in the final challenge paper. Every participant can publish their algorithms and results independently after the challenge, but we request they cite the summary paper.
The results and our corresponding evaluation of all participating teams will be made publicly available on the challenge website after the conference session.

Submission method

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

Examples:

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

Participants will create a Docker container with their algorithm, and provide this to the challenge organizers. Once the website is up, it will contain instructions on how to containerize the algorithms, and the organizers will provide support to the participants when requested. The organizers will run the Docker container on the test data set using publicly available evaluation code. The organizers will inform the participants if the Docker container fails to run, and allow the participants to provide a fix.

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.

For the validation phase: Participants are allowed to submit their solutions multiple times before the designated deadline. The evaluation will be conducted using the submitted Docker image on the validation dataset, and the performance will be displayed on the public leaderboard.

For the testing phase: No multiple submissions will be allowed. The evaluation will be performed on the latest submitted Docker image on the test dataset. Resubmissions are only allowed in cases of technical errors with Docker.

After the MICCAI challenge session, new submissions and updates may be made.

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

Training Data Release: April-May, 2024
Registration: Will be announced once the challenge is accepted
Docker Submission Deadline: 2 months before the challenge session
Top 3 teams informed: 2 weeks before the challenge, so they can prepare a presentation Complete results will be announced at the challenge session.

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).

MACC dataset: The MACC study the ethics approval from NHG DOMAIN SPECIFIC REVIEW BOARD (DSRB) APPROVAL with a reference number of 2015/00406 (approval period: 01 November 2017 to 31 March 2020). The NHG DSRB operates in accordance to the ICH GCP and all applicable laws and regulations. Subjects were recruited into the study from the NUHS memory clinics and NUHS MACC research cohorts.

SG70 dataset: The SG70 study obtained the ethics approval from NUS-IRB REVIEW BOARD APPROVAL with a reference number of 2020/398 (approval period: 28 Aug 2020 to 31 May 2022). This is human biomedical research that is regulated by the Human Biomedical Research Act (HBRA) and researchers are required by law to comply with all the relevant regulatory requirements of the HBRA. The study was conducted at NUS - Yong Loo Lin School of Medicine.

Edinburgh dataset: All primary studies that provide data for this Challenge have been approved by the corresponding ethics boards, and studies have been conducted according to the Declaration of Helsinki. The data provided and used in this challenge will be double-anonimised image data. We provide brain-extracted and co-registered in the T2-weighted native space, volumes with T1-weighted, T2-weighted and FLAIR contrast. In addition, we provide binary masks of the Regions of Interests (basal ganglia and white matter/centrum semiovale) and binary masks of the manually-edited PVS segmentations. Study IDs are also removed, and specific IDs for this Challenge are given. Data will be grouped by MRI protocol, not by primary study source. Hence, Challenge participants and the rest of the Challenge organisers will be blind to primary study characteristics.

Shanghai TCM dataset: Prospective data collection. IRB is pending approval. Data will be used for hidden test set only.

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)

Custom data license; non-commercial use only for the challenge purpose.

Terms of use:

1. When joining the EPVS Challenge users team, please make sure you provide a short description of your core team, which includes your name and academic affiliation. Without these information, we are not able to accept your request. The application link will be available on the challenge website later.

2. Access to the data requires you to agree with the following terms and conditions:
By joining the EPVS Challenge users team, you acknowledge that the owner of the EPVS Challenge Dataset available is the National University of Singapore/National University Hospital (for SG70 and MACC dataset), Longhua Hospital Shanghai University of Traditional Chinese Medicine, China (for Shanghai TCM dataset) and University of Edinburgh (for Edinburgh dataset).

3. EPVS Challenge Dataset is used for research and education only. Any other kind of use might lead to recall of all datasets, stop of collaboration and legal consequences.

4. EPVS Challenge Dataset is used for the EPVS Challenge 2024 ONLY. Any form of data sharing and distribution (including raw data and pre-processed data) is strictly not allowed. Any other kind of use might lead to recall of all datasets, stop of collaboration and legal consequences.

5. Please contact the PIs of the EPVS Challenge Dataset for formal data sharing consent if you would like to use it for any other purpose (including but not limited to publication, using this dataset in other projects).

NOTE: Information on how to download the data will be available on the challenge website upon acceptance.

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 available on the challenge website, accompanied by instructions on how participants can evaluate their methods. Participants may also opt to upload their Docker image to assess performance on the validation dataset.

Furthermore, we will provide benchmark algorithms, including standard training pipelines (e.g., U-Net), along with their performance metrics for comparison. Participants may choose to develop their algorithms using these provided baseline methods as a foundation.

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

We will encourage participating teams to make their code/submission public.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No sponsoring/funding is planned.
Only the organizers will have access to the test labels.

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

Diagnosis, Research.

Task category(ies)

State the task category(ies)

Examples:

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

Segmentation/Detection.

Cohorts

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

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

The target cohort would include healthy middle-age adults, elderly and patients with risk of dementia, have been clinically referred to a brain MRI.

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

SG70 dataset: Community-based cohort [11] where the dataset is collected for the research purpose. Healthy elderly with the risk of dementia; cognitive performance are at the mean level or slightly lower than expected for their age.

MACC: The MACC dataset is a Singaporean cohort [10] collected at the National University Hospital (NUH). It includes a significant number of participants with Mild Cognitive Impairment (MCI) and dementia, providing a diverse range of cognitive profiles for research and analysis in the context of neurodegenerative conditions.

Shanghai TCM: Community-based cohort where the dataset is collected at Longhua Hospital, Shanghai University of Traditional Chinese Medicine. Healthy elderly with the risk of dementia; cognitive performance are at the mean level or slightly lower than expected for their age.

Edinburgh dataset: There are 4 cohorts that will contribute data to this Challenge:
1) The Lothian Birth Cohort 1936 Study (https://lothian-birth-cohorts.ed.ac.uk/): community-dwelling individuals born in Edinburgh and The Lothians regions in 1936, cognitively normal, who agreed to participate in a study of cognitive aging. Image data from a brain MRI scan at a mean of 72.6 (SD 0.2) years old was used. Perivascular space burden ranges from none to moderate and confounding vascular pathology ranges from none to severe.
2) Data from a study on sleep in relation to perivascular spaces burden: participants have moderate to severe obstructive sleep apnoea (OSA), but are otherwise healthy. Mean age is 50.4 (SD 9) years old (range: 30.4 to 68.9 years). Have minimal or no vascular pathology but generally abundant perivascular space burden.
3) Observational study of Mild Stroke: patients had a mild-to-moderate ischaemic stroke, predominantly of type lacunar. Mean age: 66 years old. Scans present abundant vascular pathology and PVS burden ranging from moderate to severe.
4) STratifying Resilience and Depression Longitudinally (STRADL): Depression-focused investigation using data from Generation Scotland. Participants are healthy community-dwelling adults from 18 years old onwards. Brain MRI scans show from none to moderate vascular pathology with a PVS burden ranging also from none to moderate.

Imaging modality(ies)

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

SG70 dataset:
Magnetic Resonance Imaging: T1-weighted, T2-weighted, and FLAIR images were acquired for each subject in all three planes (with at least one image in each of the axial, sagittal, coronal planes) with high resolution (T1-weighted images: 1mm x 1mm x 1 mm; T2-weighted images: 1mm x 1mm x 3mm; T2 FLAIR images: 1mm x 1mm x 1mm).

MACC dataset:
Magnetic Resonance Imaging: T1-weighted, T2-weighted, and FLAIR images were acquired for each subject in all three planes (with at least one image in each of the axial, sagittal, coronal planes) with high resolution (T1-weighted images: 1mm x 1mm x 1 mm; T2-weighted images: 1mm x 1mm x 3mm; T2 FLAIR images: 1mm x 1mm x 3mm).

Shanghai TCM dataset: T1-weighted, T2-weighted, and FLAIR images were acquired for each subject in all three planes (with at least one image in each of the axial, sagittal, coronal planes) with high resolution (T1-weighted images: 1mm x 1mm x 5 mm; T2-weighted images: 1mm x 1mm x 5mm; T2 FLAIR images: 1mm x 1mm x 5mm).

Edinburgh dataset:
1) LBC 1936: Image data acquired in a GE Signa Horizon HDx 1.5 T clinical scanner (General Electric, Milwaukee,

WI) equipped with a self-shielding gradient set (33 mT/m maximum gradient strength) and manufacturer-supplied eight-channel phased-array head coil. The scanner was located at the Western General Hospital in Edinburgh, Scotland. T1-weighted images were acquired with 1 mm x 1 mm x 1.3 mm, T2-weighted with 1 mm x 1 mm x 2 mm, and FLAIR images with 1 mm x 1 mm x 4 mm voxel sizes. But, all images provided will have 1 mm x 1 mm x 2 mm (T2-weighted spatial resolution)
2) Studies of Sleep and Mild Stroke: Image data acquired at a 3 T (Siemens Prisma) MRI clinical scanner located at the Brain Research Imaging Centre, in the Royal Infirmary Hospital, Edinburgh, Scotland. T1-weighted and FLAIR images were acquired with voxel sizes of 1 mm3 isotropic, and the T2-weighted sequence was acquired with isotropic voxel sizes of 0.9 mm3. All images provided have been mapped into the T2-weighted space and, therefore, have voxel sizes of 0.9 mm3.
3) STRADL: Images were acquired at two places: Aberdeen and Dundee, Scotland. In Aberdeen, participants were imaged on a 3T Philips Achieva TX-series MRI system (Philips Healthcare, Best, Netherlands) with a 32 channel phased-array head coil and a back-facing mirror (software version 5.1.7; gradients with maximum amplitude 80 mT/m and maximum slew rate 100 T/m/s). In Dundee, participants were scanned using a Siemens 3T Prisma-FIT (Siemens, Erlangen, Germany) with 20 channel head and neck phased array coil and a back facing mirror (Syngo E11, gradient with max amplitude 80 mT/m and maximum slew rate 200 T/m/s). T1-weighted images were acquired with isotropic voxel sizes of 1 mm3, and the T2-weighted sequence was acquired with isotropic voxel sizes of 0.5 mm3. FLAIR images from Aberdeen have voxel sizes of 0.94 × 0.94 × 1.00 mm3, and from Dundee 1 mm3 isotropic. We provide images in the T2-weighted space with isotropic voxel sizes of 0.5 mm3.

Context information

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

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

Voxel-level segmentations of the brain:
0 - Background
1 - EPVS mask

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

No additional information.

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

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.

Enlarged perivascular spaces (EPVS) on brain MRI

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)).

Find enlarged perivascular spaces (EPVS) with high sensitivity and specificity on brain MRI.

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).

MACC Dataset: SIEMENS 3T MAGNETOM Trio + Prisma. (40 in total: 20 train; 10 validation; 10 test)

SG70 Dataset: SIEMENS 3T Prisma. (40 in total: 20 train; 10 validation; 10 test)

Shanghai TCM Dataset: Philips 3T Ingenia. (20 for testing)

Edinburgh Dataset (120 in total: 60 train; 30 validation; 30 test) :
1. LBC 1936GE Signa Horizon 1.5T; Siemens 3T Prisma. (40 in total: 20 train; 10 validation; 10 test)
2. STRADL and Studies of Sleep and Mild Stroke: Philips Achieva TX-series. (40 in total: 20 train; 10 validation; 10 test)
3. STRADL: 3T Philips Achieva TX-series; Siemens 3T Prisma-FIT. (40 in total: 20 train; 10 validation; 10 test)

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

MACC Dataset: T1-weighted Sequences: Resolution of 1mm x 1mm x 1mm. TR: 2300 ms, TE: 1.96 ms. T2-weighted Sequences: Resolution of 1mm x 1mm x 3mm. TR: 2600 ms, TE: 99 ms. T2-FLAIR: Resolution of 1mm x 1mm x 3mm. TR: 9000 ms, TE: 82 ms.

SG70 Dataset: T1-weighted Sequences: Resolution of 1mm x 1mm x 1mm, TR: 2200ms, TE: 2.45ms. T2-weighted Sequences: In-plane resolution of 1mm x 1mm x 3mm. TR: 2600, TE: 100 ms. T2-FLAIR: Resolution of 1mm x 1mm x 1mm. TR: 7000 ms, TE: 393 ms.

Shanghai TCM Dataset: T1-weighted Sequences: Resolution of 1mm x 1mm x 1mm. TR: 2300 ms, TE: 1.96 ms. T2-weighted Sequences: Resolution of 1mm x 1mm x 3mm. TR: 2600 ms, TE: 99 ms. T2-FLAIR: Resolution of 1mm x 1mm x 3mm. TR: 9000 ms, TE: 82 ms.

Edinburgh Dataset:

1. LBC 1936: T1-weighted images were acquired with 1 mm x 1 mm x 1.3 mm, T2-weighted with 1 mm x 1 mm x 2 mm, and FLAIR images with 1 mm x 1 mm x 4 mm voxel sizes. But, all images provided will have 1 mm x 1 mm x 2 mm (T2-weighted spatial resolution).
2. STRADL and Studies of Sleep and Mild Stroke: T1-weighted and FLAIR images were acquired with voxel sizes of 1 mm3 isotropic, and the T2-weighted sequence was acquired with isotropic voxel sizes of 0.9 mm3. All images provided have been mapped into the T2-weighted space and, therefore, have voxel sizes of 0.9 mm3.
3. STRADL: T1-weighted images were acquired with isotropic voxel sizes of 1 mm3, and the T2-weighted sequence was acquired with isotropic voxel sizes of 0.5 mm3. FLAIR images from Aberdeen have voxel sizes of 0.94 × 0.94 × 1.00 mm3, and from Dundee 1 mm3 isotropic. We provide images in the T2-weighted space with isotropic voxel sizes of 0.5 mm3.

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.

MACC Dataset: Memory Aging & Cognition Centre (MACC), National University Health System.

SG70 Dataset: Center for Sleep and Cognition, Yong Loo Lin School of Medicine, National University of Singapore.

Shanghai TCM dataset: Longhua Hospital Shanghai University of Traditional Chinese Medicine, China.

Edinburgh Dataset:
1. LBC 1936: Western General Hospital in Edinburgh, Scotland.
2. STRADL and Studies of Sleep and Mild Stroke: Brain Research Imaging Centre, Royal Infirmary Hospital, Edinburgh, Scotland.
3. STRADL: Aberdeen and Dundee, Scotland.

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 using clinically defined protocols.

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.

Each case in the challenge includes a 3D brain MRI of a human brain, featuring T1, T2, and T2-FLAIR sequences. We will provide participants with both raw and preprocessed images, with the preprocessing pipeline also being made available. Participants have the flexibility to either use our preprocessed images or apply their preferred preprocessing pipeline.

Training cases have an annotated label map corresponding to the EPVS.

The label maps are not provided for the test cases.

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

SG70 Dataset: Training cases: 20 3D volumes
Validation cases: 10 3D volumes
Test cases: 10 3D volumes
Total cases: 40 3D volumes

MACC Dataset: Training cases: 20 3D volumes
Validation cases: 10 3D volumes
Test cases: 10 3D volumes
Total cases: 40 3D volumes

Shanghai-TCM dataset: Test cases: 20 3D volumes

Edinburgh Dataset (120 in total: 60 train; 30 validation; 30 test):
1. LBC 1936: Training cases: 20 3D volumes; Validation cases: 10 3D volumes; Test cases: 10 3D volumes. Total cases: 40 3D volumes.
2. STRADL and Studies of Sleep and Mild Stroke: Training cases: 20 3D volumes; Validation cases: 10 3D volumes; Test cases: 10 3D volumes. Total cases: 40 3D volumes.
3. STRADL: Training cases: 20 3D volumes; Validation cases: 10 3D volumes; Test cases: 10 3D volumes. Total cases: 40 3D volumes.

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

In our EPVS segmentation challenge, the training and testing datasets are currently under preparation. The existing EPVS segmentation challenge (https://valdo.grand-challenge.org/) has demonstrated the feasibility of segmenting EPVS with a dataset comprising 6 subjects with full brain annotation for training and 10 for testing. Our multi-site, multi-cohort dataset includes 100 subjects for training, 50 for validation and 70 for independent testing, significantly increasing the data volume to enhance the robustness and accuracy of the models developed. Additionally, our dataset uniquely includes both community-based and disease cohorts. This diverse range of cases ensures that the models developed are not confined to a specific patient group but are versatile and applicable across various clinical scenarios.

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.

Our proposed dataset was collected from diverse cohorts, including both normal individuals and patients across multiple countries, thereby ensuring a broad representation of healthy and pathological anatomies as well as a rich ethnic diversity (UK, Singapore and China). This dataset, collected from various sites using different MRI scanners, introduces necessary variability in imaging conditions, reflecting the range of scenarios encountered in real-world clinical settings. Additionally, our dataset's class distribution strategy combines mirroring real-world EPVS prevalence with maintaining an equal class distribution in validation and test cases, aiming to create a balanced and realistic training and testing environment.

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.

We employed a consensus protocol approved by three trained radiologists. First, each label was created separately, with annotators A & B segmenting every second to every third slice for each label in the axial plane. Each label will be annotated by 2 annotators independently for the quality check. Disagreement will be reviewed by another independent annotator C to reach a consensus.

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.

Annotators will be trained by certified radiologists during the training phase. They will be introduced EPVS related brain anatomy and potential confusion with other cerebral vascular lesions (e.g. WMH). EPVS can be characterized on MRI as structures with dimensions less than 3mm, exhibiting small, well-defined areas of cerebrospinal fluid (CSF) intensity or closely resembling CSF intensity, and following the course of perforating vessels. (EPVS manual: https://www.ed.ac.uk/files/imports/fileManager/epvs-rating-scale-user-guide.pdf)

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

The annotators were trained radiologists with more than 5 years of experience in MRI segmentation with good inter-rater reliability.

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.

N/A

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.

T1, T2-weighted, and FLAIR MRI: Our standard preprocessing pipeline involves reorientation (to T1) and bias-field correction (all modalities) and co-registration with T1 (T2 and FLAIR). We will provide both raw and preprocessed images to the participants.

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.

Sources of error in the image annotation can originate from:
- confusion with other cerebrovascular diseases (e.g. WMH, lacune, microbleeds)
- Annotations were made mainly in the axial plane, leading to some 'noisy' labels when looking at the coronal and sagittal planes

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

None

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)

Four complementary types of performance metrics will be used to compute the rankings:
1. The Dice similarity coefficient (DSC) is used to quantify the overlap
2. The absolute volume difference(AVD) (in percentage) is used to quantify the volume difference
3. Sensitivity for individual lesions (in percentage). Individual lesions are defined as 3D-connected components (https://examples.itk.org/src/segmentation/regiongrowing/connectedcomponentsinimage/documentation).
4. Specificity for individual lesions (in percentage). Individual lesions are defined as 3D connected components (https://examples.itk.org/src/segmentation/regiongrowing/connectedcomponentsinimage/documentation).

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

We selected the Dice Similarity Coefficient (DSC) as our primary metric for the segmentation task, emphasizing its effectiveness in measuring the overlap between our predictions and the ground truth. Recognizing that Enlarged Perivascular Spaces (EPVS) are commonly evaluated in terms of volume, we have incorporated absolute volume difference(AVD) to accurately assess the volumetric aspects of our predictions. Furthermore, we have opted to use Sensitivity and Specificity for individual lesions. Our final evaluation strategy will integrate all these metrics for a comprehensive assessment of our segmentation task.

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.

We will assess the performance of algorithms in segmenting the human brain using four metrics: Dice Similarity Coefficient (DSC), Absolute Volume Difference (AVD), Sensitivity, and Specificity. For each algorithm, these metrics will be calculated across all labels in the predicted brain EPVS maps of our test set. The mean and standard deviation for each label will be determined.

Ranking will be based on performance: lower AVD scores indicate better performance, while higher scores in DSC, Sensitivity, and Specificity are preferable. The rankings for each label will be aggregated to identify the top-performing algorithm.

Finally, the results of the challenge will be run through the ChallengeR toolkit, specifically designed to calculate and display imaging challenge results.

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

The only possibility for missing data would be if an algorithm does not find any of the EPVS in the final label map. If there are missing results, the worst possible value will be used. For example, if a label does not exist in the label map, it will receive a DSC, Precision, and Recall to be 0, and the AVD will be 1 the max value of the other algorithms submitted (to ensure it is ranked last for that sub-ranking).

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

This ranking system was developed in order to take three different metric types equally into account. We also wanted to determine not just average ranking, but see if algorithms performed better in when the image was high quality vs low quality, as well as how well they performed on the pathological vs the neurotypical brains.

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.

The description of missing data handling can be found in section 'Submissions with missing results'. The mean and standard deviation of each method will be calculated using both an average of all labels as well as individually, using the ranking method as described above.

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

See above sections

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.

Inter-algorithm variability may be analyzed in the final paper, as well as an in-depth analysis as to why some algorithms performed better than others (potential problems/biases that may be present).

ADDITIONAL POINTS

References

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

[1] Wardlaw, Joanna M., et al. "Neuroimaging standards for research into small vessel disease and its contribution to aging and neurodegeneration." The Lancet Neurology 12.8 (2013) 822-838.

[2] Smith, Eric E., et al. "Harmonizing brain magnetic resonance imaging methods for vascular contributions to neurodegeneration." Alzheimer's & Dementia Diagnosis, Assessment & Disease Monitoring 11 (2019) 191-204.

[3] Doubal, Fergus N., et al. "Enlarged perivascular spaces on MRI are a feature of cerebral small vessel disease." Stroke 41.3 (2010) 450-454.

[4] Javierre-Petit, Carles, et al. "Neuropathologic and cognitive correlates of enlarged perivascular spaces in a community-based cohort of older adults." Stroke 51.9 (2020) 2825-2833.

[5] Jie, Wanxin, et al. "The relationship between enlarged perivascular spaces and cognitive function a meta-analysis of observational studies." Frontiers in Pharmacology 11 (2020) 715.

[6] Wang, Shuyue, et al. "Quantity and Morphology of Perivascular Spaces: Associations With Vascular Risk Factors and Cerebral Small Vessel Disease." J Magn Reson Imaging, Oct. 2021.

[7] Williamson, Brady J., et al. "Automated grading of enlarged perivascular spaces in clinical imaging data of an acute stroke cohort using an interpretable, 3D deep learning framework." Scientific Reports (2022).

[8] Boutinaud, Philippe, et al. "3D Segmentation of Perivascular Spaces on T1-Weighted 3 Tesla MR Images With a Convolutional Autoencoder and a U-Shaped Neural Network." Frontiers in Neuroinformatics, vol. 15, 18 June 2021.

[9] Sudre, Carole H., et al. "Where is VALDO? VAscular lesions detection and segmentation challenge at MICCAI 2021." Medical Image Analysis 91 (2024).

[10] Ji, Fang, et al. "Associations of Blood Cardiovascular Biomarkers With Brain Free Water and Its Relationship to Cognitive Decline A Diffusion-MRI Study." Neurology, vol. 101, no. 2, 11 July 2023.

[11] Xu, X., Chew, K. A., Wong, Z. X., Phua, A. K. S., Chong, E. J. Y., Teo, C. K. L., Sathe, N., Chooi, Y. C., Chia, W. P. F., Henry, C. J., Chew, E., Wang, M, et al. "The SINgapore GERiatric Intervention Study to Reduce Cognitive Decline and Physical Frailty (SINGER): Study Design and Protocol." J Prev Alzheimers Dis, vol. 9, no. 1, 2022, pp. 40-48,doi:10.14283/jpad.2022.5.

[12] Smith, Stephen M., et al. "Advances in functional and structural MR image analysis and implementation as FSL." Neuroimage, vol. 23, suppl. 1, 2004, pp. S208-S219, doi:10.1016/j.neuroimage.2004.07.051.

Further comments

Further comments from the organizers.

No