

# Ischemic Stroke Lesion Segmentation Challenge (ISLES) 2026: Structured description of the challenge design

Remark: This challenge has been slightly modified. All changes are highlighted in red.

## CHALLENGE ORGANIZATION

### Title

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

Ischemic Stroke Lesion Segmentation Challenge (ISLES) 2026

### Challenge acronym

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

ISLES'26

### 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.

Accurate infarct segmentation in brain stroke is one of the critical requirements across the disease trajectory, from acute-stage treatment guidance (e.g., reperfusion eligibility) to sub-acute and chronic-stage evaluation of patient outcome, clinical follow-up, and optimized therapeutic/rehabilitation strategies.

The Ischemic Stroke Segmentation Challenge (ISLES) has served as a foundational benchmark for the scientific community since 2015, systematically addressing infarct segmentation across diverse modalities (CT/MR), disease stages (hyper-acute to chronic), and multicenter cohorts. The established open ISLES datasets have attracted significant community interest, evidenced by over 6,000 downloads of our ISLES'22 [1] and ATLAS [2] datasets, respectively. While initial challenges focused on technical segmentation tasks (e.g., ISLES'2015 [3]), recent editions have increasingly emphasized clinical relevance and real-world transferability. For instance, ISLES'2022 algorithms were subsequently stress-tested on real-world cohorts and implemented into a standalone software for clinical utility [4]. Furthermore, our ISLES'24 challenge setting envisioned personalized patient phenotyping through detailed multimodal datasets, enabling algorithms to capture a comprehensive patient picture for final infarct prediction [5,6].

The ISLES'26 edition specifically aims to establish a robust benchmark for generalized stroke infarct segmentation on T1-weighted MRI scans, covering acute, sub-acute, and chronic disease stages. Participants will develop methodologies on the largest-ever annotated stroke infarct dataset, comprising approximately 2,000 scans sourced from over 60 international centers. This initiative directly addresses generalization gaps identified in ISLES'22-ATLAS (a prior attempt to this task we organized), by providing a 40% increase in data diversity and a greater than 2-fold expansion in training data. Furthermore, the expanded dataset also includes crucial clinical variables, enabling the organizing team to conduct post-MICCAI evaluations of algorithmic clinical utility through

analyses of patient outcome measures.

The complexity of ISLES'26 requires algorithms to robustly manage the full clinical spectrum, including diverse lesion sizes, complex anatomical patterns, and inherent heterogeneity from multi-center, multi-scanner, un-harmonized MR scans. Our ultimate goal is to facilitate the direct transfer of successful algorithms into standalone software for practical clinical utility.

For MICCAI'26, ISLES would be happy to collaborate with the SWITCH+ workshop, TopBrain2, and TopAneu to organize a joint, comprehensive full-day satellite event focusing on the advances and persistent challenges in neurovascular image analysis.

### Challenge keywords

List the primary keywords that characterize the challenge.

Stroke, brain imaging, MRI, deep learning, image segmentation, image processing

### Year

2026

### Novelty of the challenge

Briefly describe the novelty of the challenge.

Attaining a reliable stroke segmentation in T1-weighted MRI has been researched previously, including members of our team [7] and in a previous edition of our challenge (ISLES'22-ATLAS). However, to date, no available algorithm is robust enough for clinical use due to the diversity in disease timing and data variability. Consequently, automatic segmentation algorithms have failed to prove real clinical utility. As such, we are addressing this clinical-translational gap. The novelty of ISLES'26 lies, on the one hand, in enabling methodological and algorithmic developments through the largest and most diverse manually annotated T1-weighted stroke MRI dataset to date (approximately 2,000 cases). This scale and diversity are expected to foster increased robustness and generalizability, potentially enabling segmentation performance on par with that of clinical experts. On the other hand, and in contrast to prior work, ISLES'26 places emphasis on clinical relevance by extending evaluation beyond conventional segmentation metrics to include downstream clinical tasks in a post-MICCAI setting. This design allows model performance to be directly related to clinical utility. In particular, we will assess in a post-challenge setting, segmentation-derived predictions with respect to clinical (e.g. modified ranking scale, mRS) and motor (e.g. Fugl Meyer Upper Extremity [13]) outcomes, thereby providing a more meaningful and translational evaluation of algorithmic performance.

### Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

ISLES'26 targets stroke lesion segmentation in T1-weighted MRI, covering acute, sub-acute and chronic lesions.

## FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

### Workshop

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

For MICCAI'26, ISLES is willing to join the SWITCH+ workshop, TopBrain2, and TopAneu challenges to organize a joint, comprehensive full-day satellite event focusing on the advances and persistent challenges in neurovascular image analysis.

### Duration

How long does the challenge take?

2 Hours

In case you selected half or full day, please explain why you need a long slot for your challenge.

-

### 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 estimate ~ 20 participating teams based on:

- 1) The number of teams that enrolled in previous ISLES challenges, which has varied between ~10 and ~30 depending on year and task;
- 2) The large amount of database download requests received in previous ISLES editions (over 6K for our ISLES'22-ATLAS, the first attempt to the task addressed in this edition's challenge);
- 3) By its closeness with the Brain Tumor Segmentation (BraTS) challenge (in terms of anatomy, task, imaging modalities, etc.), which has received over 75 participating teams in some editions (e.g. in 2020).

We aim to advertise the challenge in existing mailing lists, social media, and send emails to previous ISLES participants. During the 10 years of the ISLES challenge, we have built a large research community that will be reached.

### Publication and future plans

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

Yes, as done in previous ISLES editions, where the benchmarking of algorithms has been published in top-ranked journals (as Nature Communications for ISLES'22 [4], Medical Image Analysis for ISLES'15 [3], Stroke for ISLES'18 [10]).

### MICCAI LNCS proceedings

Indicate if you want to offer MICCAI Springer LNCS proceedings to the participants. Publishing a proceedings volume is optional and at the discretion of each challenge's organizers. At a minimum, organizers must ensure that a description of each participant's submission is publicly available. Organizers who wish to publish MICCAI Springer LNCS proceedings must adhere to the MICCAI Satellite events publication process.

No

### Collaboration with European Society of Radiology (ESR)

In collaboration with European Society of Radiology (ESR), we announce special clinical interest topics with associated clinicians who can help with the preparation of the proposals; the best 3 challenge proposals on these topics will get the opportunity to present their challenges at the European Congress of Radiology (ECR) 2027 in a special session. If

you want to organize a challenge in collaboration with ESR on one of these topics, please reach out to the MICCAI Challenges Team ([miccai-challenges-2026@dkfz-heidelberg.de](mailto:miccai-challenges-2026@dkfz-heidelberg.de)) and we will put you in contact with the corresponding clinician.

Challenge in collaboration with ESR. Ticking 'Yes' implies that the challenge has been prepared in collaboration with the clinical contact point.

No

### **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.

ISLES is an off-site challenge hosted on [Grand-Challenge.org](https://www.grand-challenge.org). Computational resources provided by the platform for test-phase inference include a 16 GB dedicated GPU. For our presentation, we will require a projector, two microphones, and loudspeakers.

# TASK 1: ATLAS reloaded: Single channel T1-weighted stroke lesion segmentation

## SUMMARY

### Abstract

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

T1-weighted infarct segmentation in stroke is crucial to evaluate the patients' disease outcome after therapy and as a complementary marker for their clinical follow-up and long-term planning of rehabilitation. While several T1-weighted automated lesion segmentation algorithms are available, most are neither accurate nor reliable enough and require significant manual effort for quality control and correction [7,8]. As such, manual lesion segmentation remains the gold standard, despite being inefficient, subjective, and precluding the analysis of large-scale cohorts [9].

ISLES'26 addresses stroke infarct segmentation in T1-weighted MRIs through the largest annotated dataset to date, retrieved from over 60 institutions worldwide. The dataset comprises a large heterogeneity in stroke lesions and patterns, acquisition protocols and machine vendors, thus enabling the development of accurate and robust solutions. We expect ISLES'26 to close the out-of-domain generalization gap seen in prior studies.

Participants will have access to a multi-scanner and multi-vendor dataset of ~1.5k scans (~ 75% of the whole dataset) to devise automatic lesion segmentation algorithmic solutions. From a technical standpoint, the task is complex given 1) a large variability in lesion locations across diverse vascular territories, 2) heterogeneity in lesion sizes (comprising multiple small punctiform lesions, to massive lesions), 3) heterogeneity in acquired data sourcing from multiple scanners and institutions. In order to guide the development of solutions that are not only technically but also clinically relevant, participants are ranked through metrics of technical (e.g. Dice scores) and clinical impact (e.g. detection of the individual lesion instances).

We envision the outputs of ISLES'26 to inform clinical utility through extensive post-MICCAI downstream tasks evaluation. As done in prior challenges [4], we would evaluate the possibilities of translating the outperforming solutions into standalone software.

### Keywords

List the primary keywords that characterize the task.

T1-weighted MRI, stroke lesions, segmentation

## ORGANIZATION

### Organizers

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

- Benedikt Wiestler, b.wiestler@tum.de; Technical University of Munich, Germany
- Bjoern Menze, bjoern.menze@uzh.ch; University of Zurich, Switzerland
- Cansu Yalcin, cansu.yalcin@udg.edu; University of Girona, Spain
- Chris Rorden, rorden@mailbox.sc.edu, ; University of South Carolina, US

- Ezequiel de la Rosa, ezequiel.delarosa@uzh.ch; University of Zurich, Switzerland
- Jan S. Kirschke, jan.kirschke@tum.de; Technical University of Munich, Germany
- Julian Deseoe, julian.deseoe@uzh.ch; University of Zurich, Switzerland
- Mahir Khan, mhhkhan@usc.edu; University of Southern California, US
- Mariano Cabezas, mariano.cabezas@udg.edu; University of Girona, Spain
- Mauricio Reyes, mauricio.reyes@unibe.ch; University of Bern, Switzerland
- Roland Wiest, roland.wiest@insel.ch; University of Bern, Switzerland
- Ruisheng Su, r.su@tue.nl; Eindhoven University of Technology, The Netherlands
- Sook-Lei Liew, sliew@chan.usc.edu; University of Southern California, US

b) Provide information on the primary contact person.

Ezequiel de la Rosa, ezequiel.delarosa@uzh.ch; University of Zurich, Switzerland

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes- Benedikt Wiestler, Roland Wiest, Jan S. Kirschke. Challenge role:

To identify clinical needs in stroke segmentation towards supporting longitudinal patient follow-up (BW, RW, JSK).  
This supported ISLES'26 task selection.

To identify and retrieve clinical hospital data to be included in the challenge (BW, JSK).

Support the data annotation protocol (BW, RW, JSK).

Identify relevant downstream clinical tasks to evaluate models' performance in real-world clinical settings and their potential software transferability (BW, RW, JSK).

### 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

One-time event with fixed conference submission deadline

### Challenge venue and platform

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

MICCAI 2026.

b) Report the platform used to run the challenge.

grand-challenge.org

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

Yes

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

<https://atlas.grand-challenge.org/>

### Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

**No user interaction is allowed at any step.**

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 may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

**Usage of publicly available data is allowed. Teams should properly disclose their usage.**

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

**May participate but not eligible for awards.**

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

**There are no challenge prizes.**

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 whole leaderboard will be announced publicly during MICCAI'26.**

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).
- Two authors from every submission will be included as coauthors of this paper.
- A short abstract describing each algorithm will be included in the manuscript.
- Participating teams can submit their results separately without any embargo.

## 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 submit a docker image through our evaluation platform (grand-challenge). Submission instructions will be shared through our website. Besides, we will release (via Git) a docker template that participants must use to build their solutions. Under exceptional deployment failures, participants will be contacted to fix and resubmit their dockers.

Algorithms should generate two predictions: a binary mask, indicating the lesion segmentation, and its corresponding soft (probability) image. Commonly, the binary mask is a thresholded version of the probability image.

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.

There are 3 phases for this challenge:

- Train phase: Teams can evaluate the performance of their trained models by themselves. With this purpose, we will release together with the first batch of training data, a Python evaluation script that computes the performance metrics defined in this document (please check the Assessment Methods section). Evaluation scripts will be shared through GitHub. It is important to mention that there is no 'validation' set for ISLES'2026. However, participants are strongly encouraged to take validation sub-sets from the training data in order to validate their models. To promote robust model development and ensure results generalize across the heterogeneous ISLES'26 dataset, we recommend that participants utilize a stratified k-fold cross-validation strategy that accounts for institutional domain shift, lesion characteristics, and lesion maturity. Specifically, we encourage center-site stratification to prevent optimistic bias from site-specific imaging protocols, alongside lesion-size stratification to ensure models are benchmarked on challenging small infarcts rather than only larger lesions. Furthermore, participants should leverage the provided metadata for temporal stratification, utilizing the "days post stroke" column for exact timing (i.e., number of days between stroke onset and MRI acquisition) or the "chronicity" column (where 180 indicates 180 days post-stroke) for categorical timelines (records marked as "NA" indicate no information about acquisition times is available). This comprehensive validation approach ensures that local performance remains representative of the hidden test set, which features unseen institutions and a wide distribution of lesion ages and sizes. For more details about a subset of our ISLES'26 dataset, see [9].

- Sanity-check phase: Consists in a 'toy' example docker submission phase. It is solely intended for teams to test whether their devised dockers work in the remote servers. Multiple submissions to this phase are allowed.

- Test phase: Participants submit a docker that will be locally evaluated by our team over the test data. Only one submission to this phase is allowed. No evaluation or ranking will be shared until the submission system is closed by the end of the challenge. For consistency, the same evaluation scripts provided during the 'train phase' will be used for computing the different teams' performance metrics and the leaderboard



## 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
- Release of Training data (1st batch, N~1100): 1 st of April 2026
  - Release of Training data (2nd batch, N~400): 15th of May 2026
  - Opening of submission system for dockers: 15th of June 2026
  - Closing of submission system for dockers: 1st of August 2026

## 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).

All data are derived from studies that were approved by their local ethics committee and were conducted in accordance with the 1964 Declaration of Helsinki. Informed consent was obtained from all subjects. The ethics committee at the receiving site (the University of Southern California) approved the receipt and sharing of the deidentified data, which do not contain any of personal identifiers.

## 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)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

**CC-BY. Participants must agree to abide by terms of use as indicated here:**

[http://fcon\\_1000.projects.nitrc.org/indi/retro/atlas.html](http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html)

**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.

A Python script for evaluating the results will be shared together with the 1st batch of train-phase data. Code will be made available through a Github repository.

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

In order to ensure reproducibility and credibility of the algorithms, and in order to make the best-performing methods available for the research community, algorithms submitted to this challenge must use and share a working Github repository with a permissive open source license (e.g. Apache 2.0).

**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 conflicts of interest. Sook-Lei Liew, Mahir Khan and Ezequiel de la Rosa will have access to the full dataset. Jan Kirschke will have access to a subset of the dataset (N~ 200).

**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, Intervention follow up, Research, 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.

**International, multi-site retrospective cohort of acute, sub-acute and chronic stroke patients that received brain T1-weighted MR imaging and evidenced brain infarct lesions.**

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

We utilize an international, multi-site retrospective cohort of acute, sub-acute, and chronic stroke patients who received brain T1-weighted (T1w) MR imaging with evidenced ischemic lesions. While DWI is the gold standard for hyperacute/acute infarct definition (addressed in our ISLES'22 challenge [4]), ISLES'26 focuses on T1w images to ensure maximum clinical, research, and translational utility. By centering the cohort on T1w imaging, we address the entire continuum of care, from acute injury to chronic recovery, for several strategic reasons:

- Broad Availability: T1w sequences are universally acquired in large longitudinal cohorts and routine clinical follow-ups, where specialized sequences like DWI are frequently omitted.
- Translational Potential: T1w sequences are the most commonly utilized sequence in research studies of stroke brain imaging and often aim to identify biomarkers of recovery. Using T1w supports translation of findings from research to clinical practice.
- Temporal Consistency: T1w imaging maintains reliable lesion conspicuity across highly heterogeneous time points, unlike DWI, which is typically limited to the first week post-onset.
- Secondary Pathologies: T1w scans uniquely enable the quantification of long-term secondary sequelae, such as focal or global atrophy, which are critical for longitudinal outcome modeling.

- Bias Mitigation: This focus avoids the severe selection bias inherent in requiring matched multimodal (DWI/FLAIR) data, which is rarely available across all stages of the stroke recovery process.

### Imaging modality(ies)

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

T1-weighted MRI

### Context information

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

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

None.

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

None.

### Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Brain in T1-weighted MRI.

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

Brain infarcts.

### 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)).

Accuracy, Sensitivity, Specificity, Robustness, Precision

## 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).

**Data is multi-scanner and multi-center; it has been acquired on 3T and 1.5T MRI scanners over 60 institutions worldwide.**

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

**Data acquisition details for each site can be found in our publication [9].**

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.

**The data are provided from the University of Southern California, but acquired from over 20 centers/institutes worldwide. Details about the centers/institutes can be found in our publication [9]. A subset of the data used in ISLES'26 (N~600) is currently accessible from the International Neuroimaging Data-Sharing Initiative website ([http://fcon\\_1000.projects.nitrc.org/indi/retro/atlas.html](http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html)).**

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).

**Healthcare professionals operating the MRI devices in clinical routine.**

### Training and test case characteristics

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

Examples:

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

**Training and test cases represent a T1-weighted MRI scan of the brain. Training cases additionally include a voxelwise binary mask, with True labels within the infarcted area and with False labels within the non-infarcted areas. Data includes acute, subacute and chronic stroke lesions visible in T1-w.**

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

**Train set: ~1500 scans (~ 75%)**

**Test set: ~500 scans (~ 25%)**

c) How much of the data are already annotated (stratified by train test in percentage)?

**~1600 scans (~80% of the whole dataset), with ~300 annotated scans from the test-set (~60% of it)**

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

The database size is conveyed by taking into account:

- The task of the challenge (segmentation).
- The effort needed to identify and retrieve the data from the centers.
- The effort needed to manually annotate scans at a voxelwise level.

The data is split in ~75% train and ~25% test, by fulfilling the following points:

- The train and test set include a wide range of stroke lesion patterns and contain a similar mix of multisite data.
- The test set also includes data from completely new sites to examine how well the methods can generalize to external centers and cohorts which have not been included during training.
- A part of the test set (N=316) is the very same test subset used in ISLES'22. This choice enables comparing side-by-side algorithmic improvements from ISLES'22 to ISLES'26.
- It is worth noting that we do not provide a validation set. However, given the large data provided, we encourage participants to consider a subset of the training set for validation of their algorithms.

e) 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 train and test sets are comprised of similar multi-site compositions, with the test set including brand new site data to examine generalizability to an untrained center. In this way, we can evaluate performance in both scenarios, as well as performance drop-off when generalizing to new data from new centers.

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

The hidden test-set contains 500 scans, and is not available to the public (neither the MR images, nor their ground-truth) at any time.

### 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 manually segmented all lesions with a team of trained researchers per our previous protocol (see ATLAS [9] for more details). All lesions were then manually inspected for accuracy by two additional team members. Any complications were reviewed by a medical doctor with extensive neuroradiology experience.

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.

Our training protocol is described in our prior ATLAS data publication [9] : "Training for the lesion identification and tracing process involved study of in-depth neuroanatomy, standardized protocols, instructional videos, and consultations with a neuroradiologist. This protocol includes tracing the same initial set of lesions twice per person, with extensive feedback provided from multiple team members. Our standard operating procedures are freely available online (<https://github.com/npl/ATLAS/>). The training manual for ITK-SNAP [16] is freely available (<http://www.itksnap.org/docs/fullmanual.php>) and was also used as part of the lesion tracing process. For lesion

identification, each T1w MRI was opened with ITK-SNAP and examined carefully. Tracers also received training in the identification of white matter hyperintensities of presumed vascular origin [17] and perivascular spaces, which were excluded from the lesion masks as much as possible. Lesions were traced using either a mouse or stylus (i.e., Wacom Intuos Draw). All identified lesions for each subject were contained in a single image file. For lesions spanning a large number of slices (i.e., >50 slices), the “interpolation” tool was used. Upon completion, raw lesion mask files were saved and named according to a BIDS-compliant naming scheme. All files were subsequently reviewed for quality control by two additional trained team members. If changes were necessary, edits were conducted by the original tracer. Upon approval, each subject’s raw mask and T1w image were added to the raw/native space dataset, then preprocessed and added to the preprocessed dataset.”

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.

For each subject file, brain lesions were identified, and masks were manually drawn in native space. Our team identified and traced lesions using ITK-SNAP [8] (version 3.8.0). After tracing, we reviewed and edited lesion masks as necessary using a standardized quality control protocol. In a subset of the data, lesion masks were received from the originating site and edited and checked for quality by our team. All team members received lesion-tracing training and followed a standard operating protocol for tracing lesions to ensure inter-rater reliability on all manually traced masks. All lesion masks were checked for quality by two separate trained team members. During the quality control process, we ensured that the boundaries of the lesion segmentation were accurate and that all identifiable lesions in the brain were traced.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

None.

### Data pre-processing method(s)

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

To better reflect real-world data acquisition, we maintain a minimal preprocessing pipeline, providing participants the flexibility to implement their own image-processing strategies. All data and masks are released in native (raw) space. The only standardized preprocessing step is skull-stripping, performed via SynthStrip [18] for de-identification purposes.

### 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.

We previously published that our tracing protocol resulted in inter-rater reliability of  $0.76 \pm 0.14$ , and intra-rater reliability was  $0.84 \pm 0.09$  [9].

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)

The following 5 metrics will be computed for each case:

Vowel-wise metrics:

- Segmentation: Dice Similarity Coefficient.
- Volume: Absolute difference.
- Area under the Precision-Recall curve (computed over soft probabilities)

Lesion-wise metrics:

- Detection: Detection F1
- Count: Absolute difference

Metrics are defined as follows:

- Dice Similarity Coefficient (voxelwise) = Detection F1 (elementwise) =  $2TP / (2TP + FN + FP)$
- Absolute volume difference (voxelwise) = Absolute lesion count difference (elementwise) =  $|\text{Total predicted} - \text{total rater}|$ .

The lesion count in the annotated/predicted images is obtained by computing the amount of connected components per case. We will release a python script which computes all these metrics together with the training data.

We compute the area under the precision–recall curve as Average Precision (AP), i.e., the weighted mean of precision values at operating points, weighted by the increase in recall (non-interpolated step function), consistent with common implementations (e.g., scikit-learn) and the guidelines recommendations [14]. For cases with an empty ground-truth lesion mask, AP is undefined. We set AP to NaN for these cases and exclude them from the AP aggregation (mean over non-NaN cases). Lesion count differences and absolute volume differences are reported in all cases, including the empty-GT ones, which ensure models are not rewarded for spurious predictions in empty masks.

Metrics derivation and interpretation:

For ISLES'26, participants are required to submit both voxel-wise probability maps (floating-point values [0, 1]) and binary segmentation masks for each case. This dual-submission format allows for a comprehensive evaluation of both model calibration and final decision-making performance:

- Soft Probability Maps: These are used exclusively to calculate the area under the precision–recall curve using a non-interpolated step function [14]. This assesses the quality of the model's confidence across all operating points. For cases with an empty ground-truth, the metric is undefined (a NaN is assigned) and is excluded from



the ranking aggregation.

- Binary Segmentation Masks: Participants submit their best-performing binary masks by choosing their preferred cutoff and post-processing strategies. All binary-dependent metrics (i.e. the Dice Coefficient, F1-score, Lesion Count Difference, and Absolute Volume Difference) will be derived directly from these provided masks.

Handling of edge cases:

To ensure clinical reliability, ISLES'26 explicitly accounts for critical edge cases through its multi-metric evaluation design. First, the cohort spans from acute to chronic stages (e.g., >180 days post stroke onset), presenting a wide spectrum of appearances from subtle acute hypointensities to well-defined chronic encephalomalacia; benchmarking models across this temporal continuum identifies algorithms capable of distinguishing true stroke pathology from T1-mimics regardless of signal contrast. Second, for negative cases or those with extremely small lesions, we prioritize Lesion Count Difference and Absolute Volume Difference. In these scenarios, while Dice and F1-scores may only yield binary values, volumetric and count-based differences provide a more granular distinction of model performance, ensuring that algorithms are strictly penalized for spurious false-positive predictions in empty or near-empty masks. Finally, to address small embolic lesions, which carry significant prognostic weight but have a negligible impact on global Dice scores, our detection-based metrics (F1-score and Lesion Count) prioritize the identification of multi-focal edge cases, ensuring that models are not exclusively optimized for high-volume infarcts but remain sensitive to clinically critical small-scale pathology.

TP: True positives; FN: False negatives; FP: False positives.

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

The metrics considered for this task are mainly chosen from the clinical and radiological motivation of this challenge. As embolic-originated brain infarcts may have multiple small lesions which are not easy to detect, we aim to evaluate the robustness of the algorithms to localize and detect most of the present lesions. From this perspective, typical segmentation metrics (as Dice Coefficient) may not be sufficient, since small lesions may not consistently drive changes in some overlap measures (for instance, in the presence of a large stroke lesion and a very small separated embolic infarct, a large Dice increase will come by only detecting the large lesion, nonetheless the small lesion is missed).

Thus, with a focus on clinical translation, we consider metrics that are often of main interest for neuroradiologists, such as the lesion volume, the presence/absence of a lesion (i.e. detection) and the accurate count of the lesion burden. Besides, we include a classical segmentation metric as Dice Similarity Coefficient to have an overall idea of the performance overlap between ground truth and predictions. In this edition of ISLES, we further propose the inclusion of the area under the precision recall curve (PR-AUC), computed on soft (probabilistic) segmentation masks. PR-AUC provides a threshold-independent measure of overall model discrimination performance, facilitating the identification of optimal operating points and enabling explicit control over false-positive and false-negative rates. This is particularly relevant for mitigating segmentation biases arising from arbitrary threshold selection. The choice of PR-AUC (over ROC-AUC, for instance) further aims to compensate for class imbalance, where the 'healthy' brain tissue represents a much larger class than the stroke lesions. In a post-challenge setting, we additionally envisage the use of PR-AUC-guided ensembling strategies for leading solutions, using optimal thresholding and weighting schemes to derive a well-calibrated, consensus segmentation.

## 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. Ideally, provide the ranking scheme as a concrete pseudo code.

The ranking scheme considered for ISLES'2026 is the same one used in our previous ISLES editions [3,4,6,10], and consists of a 'rank then aggregate' strategy. In a nutshell, it consists of comparing each metric at the case level. Metrics are calculated for each case, followed by establishing metric-specific ranks separately for each dataset. A mean rank over all metrics is then obtained to obtain the team's rank for each case. The final teams' rank is the mean over all case-specific ranks.

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

Missing results on test cases will result in ranks for the corresponding metrics to be set to the worst possible value.

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

Robustness plays a crucial role in our challenge. The chosen ranking scheme ('rank then aggregate') proves to emphasize robustness, since it is less influenced by outliers in the case-level metrics (compared, for instance, with 'aggregate then rank' rankings). It further considers fairness and transparency for the teams. We have been using this ranking scheme since 2015, and the strategy received positive feedback from the research community and has been widely adopted for other medical image challenges.

## Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

A 1000 bootstraps ranking will be obtained using challenger [12].

T-test or Wilcoxon (for non-uniformly distributed data) will be used to check for statistically significant differences among submissions. After inspecting the data distribution, the choice of a parametric or non-parametric tests will be decided.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

Median values, which are less impacted by outliers, are reported. Confidence intervals are also reported for the median in the 1000 test-set bootstraps.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

We calculate variance and [5th, 95th] percentiles over the test set. Visualization is conducted through boxplots, which allow a complete data distribution assessment, including outliers, IQR, etc. With the analysis of variance and analysis of outliers (inspected through boxplots) we aim to get insights on the robustness of the methods.

Provide a description of how variability of rankings is assessed.

We conducted a 1000 ranking bootstraps to evaluate variability in the leaderboard. This is done using challengerR [12].

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

T-test or Wilcoxon (for non-uniformly distributed data) will be used to check for statistically significant differences among submissions. After inspecting the data distribution, the choice of a parametric or non-parametric tests will be decided.

Provide a description of the missing data handling.

Missing results on test cases will result in metrics to be set to the worst possible value. However, we rarely have missing data in our challenge, since metrics are a-priori defined to cover all possible scenarios and we further communicate and guide the teams during the test phase in case their algorithm fails in a specific subject (a common reason to have missing data).

Indicate any software product that is used for all data analysis methods.

Statistics are conducted in R.

### Further analyses

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

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

In the final challenge manuscript we will provide evaluation of algorithm-rater and inter-algorithm variability. We will conduct ensembling strategies towards identifying hybrid solutions outperforming individual methods. We will evaluate the possibility of releasing the best-performing identified algorithm (e.g. through standalone software, as we did in ISLES'22 [4]) towards clinical translation. Furthermore, we will conduct an extensive post-challenge clinical outcome analysis. While challenge participants focus strictly on lesion segmentation, we will validate model-derived lesion features against functional outcomes (mRS, Fugl-Meyer). To ensure these findings reflect true stroke-related pathology, we will perform a confounder disentanglement analysis:

- Clinical impact isolation: We will analyze the "delta mRS" (current mRS minus pre-stroke mRS) to isolate the functional deficit specifically attributable to the ischemic event from pre-existing disability.
- Structural confounder control: We will integrate patient demographics and extract structural biomarkers (e.g., brain atrophy) to quantify their influence.
- Brain age integration: Leveraging established "Brain Age" models [15], we will estimate the relative contribution of pre-existing neurodegeneration versus acute stroke pathology to the clinical outcome.
- Statistical robustness: We will report partial correlations between model-derived metrics and functional scores, explicitly adjusting for demographic and structural confounders.

These analyses will provide a transparent assessment of how much clinical predictive power is driven by the stroke lesion itself versus age-related imaging markers.

### 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] Hernandez Petzsche, Moritz R., et al. "ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset." *Scientific data* 9.1 (2022): 762.
- [2] Liew, Sook-Lei, et al. "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms." *Scientific data* 9.1 (2022): 320.
- [3] Maier, Oskar, et al. "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI." *Medical image analysis* 35 (2017): 250-269.
- [4] de la Rosa, Ezequiel, et al. "DeepISLES: a clinically validated ischemic stroke segmentation model from the ISLES'22 challenge." *Nature Communications* 16.1 (2025): 7357.
- [5] Riedel, E. O., et al. "ISLES'24—A Real-World Longitudinal Multimodal Stroke Dataset. arXiv 2024." arXiv preprint arXiv:2408.11142. Under revision at *Radiology: AI*.
- [6] de la Rosa, Ezequiel, et al. "ISLES'24: Improving final infarct prediction in ischemic stroke using multimodal imaging and clinical data." arXiv preprint arXiv:2408.10966 (2024). Under revision at *Medical Image Analysis*.
- [7] Liew, S.-L. et al. The ENIGMA Stroke Recovery Working Group: Big data neuroimaging to study brain-behavior relationships after stroke. *Human brain mapping*
- [8] Ito, K. L., Kim, H. & Liew, S. L. A comparison of automated lesion segmentation approaches for chronic stroke T1weighted MRI data. *Human brain mapping* 40, 4669–4685 (2019).
- [9] Liew, Sook-Lei, et al. "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms." *Scientific data* 9.1 (2022): 320.
- [10] Hakim, Arsany, et al. "Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: Lessons from the ISLES challenge." *Stroke* 52.7 (2021): 2328-2337.
- [11] [https://surfer.nmr.mgh.harvard.edu/fswiki/mri\\_deface](https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface)
- [12] Wiesenfarth, M. et al. Methods and open-source toolkit for analyzing and visualizing challenge results. *Sci. Rep.* 11, 1–15 (2021).
- [13] Liew, Sook-Lei, et al. "Association of brain age, lesion volume, and functional outcome in patients with stroke." *Neurology* 100.20 (2023): e2103-e2113.
- [14] Maier-Hein, Lena, et al. "Metrics reloaded: recommendations for image analysis validation." *Nature methods* 21.2 (2024): 195-212.]
- [15] MarinPardo, Octavio, et al. "Brain age is longitudinally associated with sensorimotor impairment and mild cognitive impairment in subacute stroke." *Journal of the American Heart Association* 14.20 (2025): e041603.
- [16] Yushkevich, P. A. & Gerig, G. ITK-SNAP: an interactive medical image segmentation tool to meet the need for expert-guided segmentation of complex medical images. *IEEE pulse* 8, 54–57 (2017).
- [17] Wardlaw, J. M. et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology* 12, 822–838 (2013).
- [18] Hoopes, Andrew, et al. "SynthStrip: skull-stripping for any brain image." *NeuroImage* 260 (2022): 119474.

## Further comments

Further comments from the organizers.

Several members of the organization committee have participated in the organization of other MICCAI challenges (such as Brats, TopBrain, TopCow, Verse, among others). We finally would like to thank the reviewers for taking the time to evaluate our ISLES 2026 proposal.