

BraTS Generalizability Across Tumors: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

BraTS Generalizability Across Tumors

Challenge acronym

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

N/A

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.

The International Brain Tumor Segmentation (BraTS) challenge has been focusing, since its inception in 2012, on the generation of a benchmarking environment and a dataset for the delineation of adult brain gliomas. The focus of BraTS2023 challenge remained the same in terms of generating the common benchmark environment, while the dataset expands into explicitly addressing 1) the same adult glioma population, as well as 2) the underserved sub-Saharan African brain glioma patient population, 3) brain/intracranial meningioma, 4) brain metastasis, and 5) pediatric brain tumor patients. Although segmentation is the most widely investigated medical image processing task, the various challenges have been organized to focus only on specific clinical tasks. That is, each segmentation method was evaluated exclusively on the patients population it was trained on in each sub-challenge. In this challenge, we aim to organize the Generalizability Assessment of Segmentation Algorithms Across Brain Tumors. The hypothesis is that a method capable of performing well on multiple segmentation tasks will generalize well on unseen tasks. Specifically, in this task, we will be focusing on assessing the algorithmic generalizability beyond each individual patient population and focus across all of them. Importantly, although each MR exams will undergo the same preprocessing pipeline, including an intensity normalization step, there are characteristics of each exam that will not be affected (i.e., different number of lesions per exam, different location within the brain, etc.) preserving the generalizability aspect of the challenge.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Generalizability, Brain Tumors, Segmentation, Cancer, Challenge, Glioma, Glioblastoma, dipg, Metastases, Meningioma, NIH, DREAM, diffuse glioma

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.

/

Publication and future plans

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

/

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.

N/A

TASK 1: BraTS-GoAT

SUMMARY

Abstract

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

The International Brain Tumor Segmentation (BraTS) challenge has been focusing, since its inception in 2012, on the generation of a benchmarking environment and a dataset for the delineation of adult brain gliomas. The focus of BraTS2023 challenge remained the same in terms of generating the common benchmark environment, while the dataset expands into explicitly addressing 1) the same adult glioma population, as well as 2) the underserved sub-Saharan African brain glioma patient population, 3) brain/intracranial meningioma, 4) brain metastasis, and 5) pediatric brain tumor patients. Although segmentation is the most widely investigated medical image processing task, the various challenges have been organized to focus only on specific clinical tasks. That is, each segmentation method was evaluated exclusively on the patients population it was trained on in each sub-challenge. In this challenge, we aim to organize the Generalizability Assessment of Segmentation Algorithms Across Brain Tumors. The hypothesis is that a method capable of performing well on multiple segmentation tasks will generalize well on unseen tasks. Specifically, in this task, we will be focusing on assessing the algorithmic generalizability beyond each individual patient population and focus across all of them. Importantly, although each MR exams will undergo the same preprocessing pipeline, including an intensity normalization step, there are characteristics of each exam that will not be affected (i.e., different number of lesions per exam, different location within the brain, etc.) preserving the generalizability aspect of the challenge.

Keywords

List the primary keywords that characterize the task.

Meningioma, NIH, DREAM, diffuse glioma

ORGANIZATION

Organizers

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

Lead Organizers

Gian Marco Conte, Mayo Clinic, USA

Ujjwal Baid, Indiana University

Spyridon Bakas, Indiana University

Associate organizing committee (alphabetical)

Mariam Aboian, Yale University

Maruf Adewole, Crestview Radiology Ltd., Nigeria

Jake Albrecht, Sage Bionetworks

Udunna Anazodo, McGill University, Canada/Crestview Radiology Ltd., Nigeria

Evan Calabrese, Duke Center for Artificial Intelligence in Radiology, Duke University Medical Center

Verena Chung, Sage Bionetworks

Anastasia Janas, Yale University

Anahita Fathi Kazerooni, Children's Hospital of Philadelphia/University of Pennsylvania

Dominic Labella, Duke University Medical Center

Marius George Linguraru, Children's National Hospital/George Washington University

Bjoern Menze, University of Zurich, Switzerland

Ahmed Moawad, Mercy Catholic Medical Center

Jeffrey Rudie, Scripps Clinic and University of California, San Diego
Taki Shinohara University of Pennsylvania, Philadelphia, PA, USA

b) Provide information on the primary contact person.

Conte.gianmarco@mayo.edu

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 submission deadline.

Challenge venue and platform

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

ISBI2024

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

Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge [1], we have coordinated with them and following the support from NIH (represented by Dr Keyvan Farahani in the organizing committee - Chair of the NCI AI Challenges Working Group) Synapse will be used as the platform to drive the evaluation of this challenge.

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

www.synapse.org/brats_goat

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.

Participants are not allowed to use additional data neither from publicly available datasets nor their own institutions.

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 organizers and their immediate groups will not be eligible for awards. Since organizing institutions are large, other employees from other labs/departments may participate and should be eligible for the awards and to be listed in the leaderboard.

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

We will provide plaques to the top 3 performing teams. NIH will also provide Certificates of Merit to the top 3 performing teams.

Awards: no monetary awards will be available.

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.

Top 3 performing methods will be announced publicly at the conference and the participants will be invited to present their method.

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).

The participants will be invited to publish their methods in the ISBI IEEE Xplore Challenge post-conference proceedings. Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of the challenge.

Submission method

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

Examples:

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

The participants are required to send the output of their methods to the evaluation platform for the scoring to occur during the training and the validation phases. At the end of the validation phase the participants are asked to identify the method they would like to evaluate in the final testing/ranking phase. The organizers will then confirm receiving the containerized method and will evaluate it in the hidden testing data. The participants will be provided guidelines on the form of the container as we have done in previous years. This will enable confirmation of reproducibility, running of these algorithms to the previous BraTS instances and comparison with results obtained by algorithms of previous years, thereby maximizing solutions in solving the problem of brain tumor segmentation. During the training and validation phases, the participants will have the chance to test the functionality of their submission through both the Cancer Imaging Phenomics Toolkit (CaPTk [5-6], <https://github.com/CBICA/CaPTk>), and the Federated Tumor Segmentation (FeTS) Tool [7] (<https://fetsai.github.io/Front-End/>) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

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.

We intend to release the validation set in January together with the training set, allowing participants to tune their methods in the unseen validation data. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform will be allowed for the validation phase. Only 2 submissions will be allowed in the final testing/ranking data/phase.

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

Registration dates: From now until submission deadline of short papers reporting method and preliminary results (see below).

Jan 8: Website will be available

Jan 15: Availability of complete training data (with ground truth labels) and validation data (without ground truth labels).

Mar 15: Submission of short papers reporting method & preliminary results.

Mar 30: Submission of containerized algorithm to the evaluation platform.

April 10: Evaluation on testing data (by the organizers - only for participants with submitted papers).

April 20: Contacting top performing methods for preparing slides for oral presentation.

May 27-30: Announcement of final top 3 ranked teams: Challenge at ISBI

Jul 15: Camera-ready submission of extended papers

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).

We are already in close coordination with The Cancer Imaging Archive (TCIA) and the Imaging Data Commons (IDC) of the National Institutes of Health (NIH), to release the training and validation data following their standard licensing (<https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions>)

The TCIA has already approved this, and we are now in the process of submission (includes a detailed curation process specific to TCIA). The cloud-based IDC is routinely updated with new collections from TCIA. IDC public collections are now part of the Google Public Datasets Program. This will effectively make all the BraTS data available in the Google Marketplace, increasing the potential for access to the data and downstream AI developments using Google's AI resources. IDC data are also expected to be available through the AWS (Amazon Web Services) Marketplace.

Informed consent was obtained from all subjects at their respective institutions, and the protocol for releasing the data was approved by the institutional review board of the data-contributing institution.

Data usage agreement

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

Examples:

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

CC BY.

Additional comments: CC-BY, but if any of the non-TCIA contributors object to this license, the specific subset of the BraTS data will be released under a CC-BY-NC license.

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 preprocessing tools, evaluation metrics, and the ranking code used during the whole challenge's lifecycle will be made available through the Cancer Imaging Phenomics Toolkit (CaPTk [5-6], <https://github.com/CBICA/CaPTk>), and the Federated Tumor Segmentation (FeTS) Platform [7] (<https://fets-ai.github.io/Front-End/>).

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

The participants are required to submit their containerized algorithm, during or after the validation phase. Specific instructions for the containerization will be provided after the challenge approval. These instructions will be very similar to what we were requesting participants to provide during the BraTS2023 challenge.

These submitted containers will be made publicly available to the scientific community.

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.

Ujjwal Baid, Spyridon Bakas, SAGE Bionetworks, the clinical evaluators, will have access to the validation, and test case 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

Treatment planning, Intervention planning, Assistance, Research, Surgery, Training, Diagnosis, CAD, Education, Decision support.

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.

Retrospective multi-institutional cohort of patients, diagnosed with a brain tumor, clinically scanned with mpMRI acquisition protocol including i) pre-contrast and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Retrospective multi-institutional cohort of patients, diagnosed with brain tumor, clinically scanned with mpMRI acquisition protocol including i) pre-contrast and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Imaging modality(ies)

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

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).

Directly to the image data (i.e., tumor sub-region volumes)

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

N/A

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 mpMRI scans.

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.

Tumor in the brain.

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)).

Lesion-wise Dice and Hausdorff 95th percentile

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).

The exact scanners and their technical specifications used for acquiring the TCIA cohort has been listed in the data reference published in our related manuscripts [1,2,4]. Since then, multiple institutions have contributed data to create the current RSNA-ASNR-MICCAI BraTS dataset and these are listed in the latest BraTS arxiv paper [1]. We are currently in coordination with TCIA to make the complete BraTS 2021-2023 dataset permanently available

through their portal. All the acquisition details will be included together with the data availability in TCIA, and subsequently in IDC.

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

The acquisition protocols are different across (and within each) contributing institution, as these represent scans of real routine clinical practice. Specific details (e.g., echo time, repetition time, original acquisition plane) of each scan of each patient will be published as supplementary material together with the challenge meta-analysis manuscript when available.

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 provided data describe mpMRI scans, acquired with different clinical protocols and various scanners as described in the previous tasks.

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).

People involved in MRI acquisition for suspected and diagnosis of brain tumors during standard clinical practice.

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.

A case describes multi-parametric MRI scans for a single patient at a single timepoint. The exact scans included for one case are i) unenhanced and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Please note that all sequences included for each case of the provided dataset, represent the sequences with the best image quality available in the acquiring institution for this particular case. There was no inclusion/exclusion criterion applied that related to 3d acquisitions, or the exact type of pulse sequence (for example MPRAGE). We, instead, accepted all types of T1 acquisitions (with the exception of T1 FLAIR, as we did not want to mix the fluid suppressed values with non-flair scans) and then we applied the harmonized preprocessing protocol we have been using in BraTS, across the complete data. This preprocessing ensures all scans have 3D representations on a specific resolution (1mm³), and aligned to the same anatomical atlas.

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

Training data: 2251 cases

Validation data: 319 cases

Testing data: 195. We are also working on extending the test set.

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

Based on availability. The data was split in these numbers between training, validation, and testing after considering the number of cases used as test cases in previous instances of BraTS and the fact that the organizers did not want to reveal ground truth labels of previous test cases, to avoid compromising ranking the participants.

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.

N/A

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.

Automated segmentations were generated using the FeTS tool and annotators corrected them to create ground truth dataset. These annotations were approved by radiologists.

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

The data considered in this task of the BraTS 2023 challenge follows the paradigm of the BraTS 2021-2022 challenge data. The annotation of these data followed a pre-defined clinically-approved annotation protocol (defined by expert neuroradiologists), which was provided to all clinical annotators, describing in detail instructions on what the segmentations of each tumor sub-region should describe (see below for the summary of the specific instructions). The annotators were given the flexibility to use their tool of preference for making the annotations, and also follow either a complete manual annotation approach, or a hybrid approach where an automated approach is used to produce some initial annotations followed by their manual refinements.

Summary of specific instructions:

- i) the enhancing tumor (when present) delineates the hyperintense signal of the T1-Gd, after excluding the vessels.
- ii) the necrotic core (when present) outlines regions appearing dark in both T1 and T1-Gd images (denoting necrosis/cysts), and darked regions in T1-Gd that appear brighter in T1.
- iii) the tumor core, which is the union of the enhancing tumor and the necrotic core described in (i) and (ii) above.
- iv) the farthest tumor extent including the edema (what is called the whole tumor), delineates the tissue represented by the abnormal T2-FLAIR envelope.

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.

Each case was assigned to a pair of annotator-approver. Annotators spanned across various experience levels and clinical/academic ranks, while the approvers were the 2 experienced board-certified neuroradiologists (with >15 years of experience), listed in the "Organizers" section as "clinical evaluators and annotation approvers". The annotators were given the flexibility to use their tool of preference for making the annotations, and also follow either a complete manual annotation approach, or a hybrid approach where an automated approach is used to produce some initial annotations followed by their manual refinements. Once the annotators were satisfied with the produced annotations, they were passing these to the corresponding approver. The approver is then responsible for signing off these annotations. Specifically, the approver would review the tumor annotations, in tandem with the corresponding MRI scans, and if the annotations were not of satisfactory quality they would be sent back to the annotators for further refinements. This iterative approach was followed for all cases, until their respective annotations reached satisfactory quality (according to the approver) for being publicly available and noted as final ground truth segmentation labels for these scans.

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.

No Aggregation

Data pre-processing method(s)

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

The exact preprocessing pipeline applied to all the data considered in the BraTS 2024 challenge is identical with the one evaluated and followed by the BraTS 2017-2023 challenges. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format [10], we first perform a re-orientation of all input scans (T1, T1- Gd, T2, T2-FLAIR) to the LPS/RAI orientation, and then register all of them to the same anatomical atlas (i.e., SRI-24 [9]) and interpolating to the same resolution as this atlas (1 mm³). The exact registration process comprises the following steps:

STEP 1: N4 Bias field correction (notably the application of N4 bias field correction is a temporary step. Taking into consideration we have previously [4] shown that use of non-parametric, non-uniform intensity normalization (i.e., N4) to correct for intensity non-uniformities caused by the inhomogeneity of the scanner's magnetic field during image acquisition obliterates the MRI signal relating to the abnormal/tumor regions, we intentionally use N4 bias field correction in the preprocessing pipeline to facilitate a more optimal rigid registration across the different MRI sequences. However, after obtaining the related information (i.e., transformation matrices), we discard the bias field corrected scans, and we apply this transformation matrix towards the final co-registered output images used in the challenge).

STEP 2: Rigid Registration of T1, T2, T2-FLAIR to the T1-Gd scan, and obtain the corresponding transformation matrix.

STEP 3: Rigid Registration of T1-Gd scan to the SRI-24 atlas [9], and obtain the corresponding transformation matrix.

STEP 4: Join the obtained transformation matrices and applying aggregated transformation to the LPS-oriented scans.

STEP 5: After completion of the registration process, we perform brain extraction to remove any apparent nonbrain tissue (e.g., neck fat, skull, eyeballs) based on a deep-learning approach we developed in house, focusing on scans with apparent brain tumors and exhaustively evaluated it in both private and public multi-institutional data [11]. We then manually assessed all scans for confirming the correct brain extraction (i.e., skull stripping), where the complete brain region is included, and all non-brain tissue is excluded. This whole pipeline, and its source code are available through the CaPTk [5-6](<https://github.com/CBICA/CaPTk>) and FeTS [7] (<https://fetsai.github.io/Front-End/>) platforms.

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.

Study and evaluation of the effect of this error is addressed by the uncertainty task of BraTS 2019-2020 (i.e., to quantify the uncertainty in the tumor segmentations) [8] and is outside the scope of the BraTS 2022 challenge.

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

N/A

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Dice Similarity Coefficient (DSC),

95% Hausdorff distance (HD)

The regions evaluated using these metrics describe the whole tumor, the tumor core, and the enhancing tumor (when present). Note that the tumor core includes the part of the tumor that is typically resected (i.e., enhancing, non-enhancing, and necrotic tumor), and the whole tumor describes all tumor sub-regions (i.e., tumor core and edema/invasion).

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

In terms of the assessed and evaluated tumor sub-regions:

- i) the enhancing tumor describes the regions of active tumor and based on this, clinical practice characterizes the extent of resection.
- ii) the tumor core (incl. the necrotic component) describes what is typically resected during a surgical procedure.
- iii) the whole tumor as it defines the whole extent of the tumor, including the peritumoral edematous tissue and highly infiltrated area.

In terms of evaluation metrics, we use:

- i) the Dice Similarity Coefficient, which is commonly used in the assessment of segmentation performance,
- ii) the 95% Hausdorff distance as opposed to standard HD, in order to avoid outliers having too much weight,
- iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or undersegment.
- iv) Precision to complement the metric of Sensitivity (also known as recall).

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.

For ranking of multidimensional outcomes (or metrics), for each team, we will compute the summation of their ranks across the average of the metrics described above as a univariate overall summary measure. This measure will decide the overall ranking for each specific team.

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

If an algorithm fails to produce a result metric for a specific test case, this metric will be set to its worst possible value (0 for the DSC and the image diagonal for the HD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Shinohara), and also while considering transparency and fairness to the participants.

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.

Similar to BraTS 2017-2023, uncertainties in rankings will be assessed using permutational analyses [2].

Performance for the segmentation task will be assessed based on relative performance of each team on each tumor tissue class and for each segmentation measure. These will be combined by averaging ranks for the measures, and statistical significance will be evaluated only for the segmentation performance measures and will be quantified by permuting the relative ranks for each segmentation measure and tissue class per subject of the testing data.

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

This permutation testing would reflect differences in performance that exceeded those that might be expected by chance.

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.

N/A

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.

N/A

Further comments

Further comments from the organizers.

N/A