The Brain Tumor Segmentation (BraTS) Cluster of Challenges: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

The Brain Tumor Segmentation (BraTS) Cluster of Challenges

Challenge acronym

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

BraTS Cluster of Challenges

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 Brain Tumor Segmentation (BraTS) Cluster of Challenges 2025 is a collaborative effort with the "AI for Response Assessment in Neuro-Oncology" (AI-RANO) cooperative group and leading clinical societies, including RSNA, ASNR, and ESNR. This partnership aims to establish and promote clinically relevant challenges to maximize the potential clinical impact of innovative algorithmic contributions from participating teams.

Since its inception at MICCAI 2012, BraTS has advanced brain glioma segmentation by enabling and benchmarking algorithmic developments, providing high-quality annotated datasets. In 2023, BraTS expanded to include multiple benchmarks, quantifying tumors beyond glioma (brain metastases, meningiomas, pediatric brain tumors, and sub-Saharan patient populations), histology images, and image synthesis. This novel, innovative design creates a benchmarking ecosystem for the systematic comparison of algorithms across diverse tasks and clinical challenges.

The significance of BraTS 2025 lies in its focus on addressing actual clinical needs across brain tumors, spanning i) tumor entities (for which there is currently limited publicly available annotated data), ii) disease course (pre- and post-treatment, for reliable longitudinal assessment of tumor response), iii) domains (radiology & histopathology), & iv) computational tasks (e.g., segmentation, synthesis). Corroborating BraTS's goal to address clinical needs, authoritative and leading federal and clinical organizations have partnered with BraTS (such organizations include NIH, FDA, RANO, RSNA, ASNR, ASFNR, and CBTN).

In 2025, the BraTS Cluster of Challenges will encompass 12 tasks, designed to benchmark and advance the current state-of-the-art for addressing (Task 1) Pre- and Post-Treatment Adult Glioma, (Task 2) Pre-Treatment and (Task 3) Pre-RT intracranial Meningioma, (Task 4) Pre- and Post-Treatment Brain Metastases, (Task 5) Brain Glioma in the underserved sub-Saharan African patient population, (Task 6) Pre-Treatment Pediatric Tumor Patients in partnership with multiple related societies, (Task 7) Generalizability of Segmentation Methods Across Tumors, (Task 8) Evaluation of Augmentation Techniques, in partnership with FDA, (Task 9) MRI Synthesis, (Task 10) MRI

Inpainting, (Task 11) Assessing the Heterogeneous Histologic Landscape of Glioma, as well as (Task 12) Predicting the Tumor Response During Therapy. Detailed descriptions of each task are provided in the following sections. BraTS 2025 participants can obtain the training and validation data of the challenge at any point from the Synapse platform. These datasets will be used to develop, containerize, and finally evaluate their algorithms in unseen validation data until August 2025, when the organizers will stop accepting new submissions and evaluate the submitted algorithms in the hidden testing data. To ensure accurate performance evaluation, expert neuroradiologists, and neuropathologists create and approve ground truth annotations for all datasets for each subject in the training, validation, and testing datasets. Notably, in 2025, multiple raters will independently annotate a subset of test cases for several Tasks to compare algorithmic performance against human expert inter-rater variability.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

BraTS, Brain Tumor, Neuro-Oncology, Radiology, Pathology, Digital Pathology, Pre-treatment, Post-treatment, Segmentation, Classification, Generalizability, Augmentation, Synthesis, Inpainting, Infill, Glioma, Glioblastoma, Brain metastasis, Meningioma, MRI, Sub-Saharan Africa, Pediatric, NIH, MICCAI, NCI, DREAM, RSNA, ASNR, Precision, FDA, Synapse, RANO, Progression

Year

2025

Lighthouse challenge agreement

The organizers agree to all of the following points:

- The full labeling protocol will be sent to the challenge chairs in addition to the full proposal document.
- A set of a few representative data samples including annotations will be sent to the challenge chairs in addition to the full proposal document.
- The challenge will be open for at least 4 months.
- For the dataset review, the challenge chairs will get access to the data at least 3 months before challenge opening.

Challenge organizers have read and agree to all of the above terms and conditions.

Lighthouse challenge information

In two sentences or less, what sets your challenge apart from ordinary MICCAI challenges. In other words: What makes your challenge a lighthouse challenge?

Since its inception at MICCAI 2012, the BraTS challenge has advanced the state-of-the-art in brain glioma image analysis by benchmarking algorithmic developments, providing high-quality annotated datasets, and fostering collaboration among researchers. The BraTS 2025 challenge stands out for its comprehensive approach, addressing real-world clinical needs across various brain tumor entities, disease stages, imaging modalities, and computational tasks, including segmentation, synthesis, and longitudinal assessment of tumor response.

Previous challenge(s)

What is the closest challenge to your proposed lighthouse challenge? Are there previous versions of it? Specifically, if you applied for a 2024 challenge, what is the delta between the two iterations? (e.g., number of centers for new data, number of newly added data – This is not to be confused with details about the total data set)

The 2025 BraTS Cluster of Challenges expands upon previous iterations in two key ways:
i) To highlight clinical relevance, the challenge will integrate assessment across the full disease course: Algorithms will be evaluated not only on pre-treatment scans, but also on their ability to analyze post-treatment imaging data. This longitudinal perspective provides a more comprehensive view of algorithm performance, and is complemented by BraTS-Pro (Task 12), which specifically asks participants to predict brain tumor progression, a addressing a crucial clinical need.
ii) A subset of the test cases will undergo an independent re-annotation process to assess inter-rater variability. This allows the challenge to gauge how well algorithms perform against rigorously validated ground truth data - a novel aspect that has not been addressed in prior BraTS challenges. By scrutinizing the ground truth itself, the 2025 challenge sets a new standard for robustness in evaluating state-of-the-art methods.
The multi-timepoint assessment and meticulous ground truth validation make the 2025 BraTS Challenge the most thorough evaluation of brain tumor algorithms to date. Participants can expect rankings that truly reflect real-world performance across the entire disease trajectory.

Test set status

Was the test set (or parts of it) already used in previous challenges and/or previously made publicly available?

Parts of the test set have been used in prior iterations, but were never made publicly available. In addition, we will add additional test cases (and subtasks) as well as inter-rater annotations across the spectrum of BraTS2025.

What major scientific advances or insights are expected from the challenge?

Please describe the major scientific advances ore insights you expect to be gained from the challenge. Please include references to the state of the art in your description and list open research questions to which the challenge seeks answers or solutions.

The clinical management of patients with brain tumors remains a significant challenge. However, advances in innovative image analysis methods promise to better support clinical decision-making, such as therapy planning and response assessment. The BraTS 2025 Cluster of Challenges is a major driving force for algorithm development in neuro-oncology, poised to significantly enhance our understanding and management of brain tumors across various tumor types, imaging modalities, and stages of the disease course.
By leveraging a multifaceted approach, BraTS 2025 aims to foster the development and clinical translation of cutting-edge algorithms, thereby driving innovation in medical imaging and tumor characterization. One of the key strengths of BraTS 2025 lies in its comprehensive and collaborative framework, which brings together a diverse array of scientific and clinical expertise. This initiative is endorsed by leading clinical societies, underscoring its clinical relevance and potential impact on patient care.
Participants are tasked with addressing a spectrum of challenges across the spectrum of brain tumors. Importantly, this close collaboration ensures that the tasks outlined in the BraTS Cluster of Challenges address unmet clinical needs. By fostering the development of innovative solutions, BraTS 2025 has the potential to significantly advance the field of neuro-oncology.

Clinical body affiliation

Please describe your proposed challenge's affiliation with a clinical body, if any. How do you plan to engage the clinical community that your challenge is set to impact?

Corroborating the BraTS goal to address clinical needs, authoritative and leading federal and clinical organizations have partnered with BraTS (such organizations include NIH, FDA, RANO, RSNA, ASNR, ASFNR, and CBTN). BraTS has a long-standing tradition of intensely engaging with the clinical community, highlighted by the "RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification", which was jointly organized by leading organizations in the clinical and computer science communities.

Deadline for data acquisition and annotation

What's the deadline for data acquisition and annotation?

For the 2025 BraTS Cluster of Challenges, we plan to finish data acquisition and annotation by the end of 2024.

How much prize money has been secured?

Please state how much prize money has already been secured for the challenge.

We are currently in discussion with potential sponsors for the BraTS 2025 Lighthouse Challenge. Please note that Intel has been offering monetary awards for each of BraTS since 2018, Neosoma for BraTS 2021-2022, and Cortechs.ai during 2023. NIH will also provide Certificates of Merit to the top 3 performing teams.

Computing requirements per participant

Roughly estimate how much computing power would be required per challenge participant?

This is depending on the task chosen to be addressed by a participant. For most tasks, well-available computing/GPU hardware already allows the development and training of algorithms.

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

The BraTS 2025 Cluster of Challenge is part of the BrainLes workshop at MICCAI.

Duration

How long does the challenge take?

Full day.

Expected number of participants

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

We can conservatively estimate approximately 1,000 participating teams for this year's cluster of challenges. This is based on the continuously increasing number of teams participating in the BraTS challenge during its initial 10 years (2012:n=10, 2013: n=10, 2014: n=10, 2015: n=12, 2016: n=19, 2017: n=53, 2018: n=63, 2019: n=72, 2020: n=78, 2021: n>1,000). Notably, we strongly believe that the 2021 participation increase is a result of multiple factors (challenge maturity, involvement of RSNA & ASNR, professional evaluation through Synapse and Kaggle)

that has been carried forward since 2021 up to this year's cluster of challenges, thereby guaranteeing broad participation.
We also advertise the event in related mailing lists (e.g., CVML; visionlist@visionscience.com; cvnet@mail.ewind.com; MIPS@LISTSERV.CC.EMORY.EDU), NCI's CBIIT blog posts and tweets, and we intend to send an email to all the above and notify them about this year's challenge.
Finally, since we will specifically focus on assessing generalizability across brain tumors and patient populations (including the sub-Saharan African population), we will also advertise the event in ML communities in Africa to strengthen local participation. Communities we will consider include the "Data Science Nigeria" (DSN, https://www.datasciencenigeria.org), the "African Institute of Mathematical Sciences" (https://aims.edu.gh), the "African Centre of Excellence in Data Science" (ACE-DS, https://aceds.ur.ac.rw/), and the "INDABA" (https://deeplearningindaba.com).

Publication and future plans

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

We intend to coordinate 2 specific publication plans immediately after the challenge.
Plan 1:
Coordination of the BraTS proceedings with the BrainLes proceedings allows the BraTS participants to publish their methods in the associated Springer LNCS post-conference proceedings. We have already been doing this for BraTS since 2015.
Plan 2:
We will coordinate journal manuscripts focusing on publishing and summarizing the results of each BraTS 2025 challenge, making a comprehensive meta-analysis for each to inform the community about the obtained results, findings, and insights.

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.

Hardware requirements for the in-person meeting: 1 projector, 3 microphones, loudspeakers
BraTS 2025 Cluster of Challenges describes off-site challenges, where 1) during the training phase, algorithms are trained using the participants' computing infrastructure, and 2) during the validation and final testing/ranking phase using the organizers' infrastructure (i.e., Synapse.org - SAGE Bionetworks).

TASK 1: BraTS-Glioma: Glioma Segmentation on Pre- and Post-treatment MRI

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.

Brain tumors are among the deadliest types of cancer. Specifically, glioblastoma, and diffuse astrocytic glioma with molecular features of glioblastoma (WHO Grade 4 astrocytoma), are the most common and aggressive malignant primary tumors of the central nervous system in adults, with extreme intrinsic heterogeneity in appearance, shape, and histology, with a median survival of approximately 12 months. Brain tumors in general are challenging to diagnose, hard to treat, and inherently resistant to conventional therapy because of the challenges in delivering drugs to the brain. Years of extensive research to improve diagnosis, characterization, and treatment have decreased mortality rates in the U.S. by 7% over the past 30 years. Although modest, these research innovations have not translated to improvements in survival.

Considering the clinical impact, the BraTS 2025 Glioma challenge will differ from previous years by combining the BraTS 2023 and 2024 datasets into a new dataset that includes both pre- and post-operative diffuse gliomas. This dataset consists of routine clinically acquired, multi-institutional multiparametric magnetic resonance imaging (mpMRI) scans of brain glioma patients. These data will be used by participants to develop, containerize, and evaluate their algorithms on unseen validation data until July 2025, when the organizers will stop accepting new submissions and evaluate the submitted algorithms on the hidden testing data. Ground truth reference annotations for all datasets are created and approved by expert neuroradiologists for every subject included in the training, validation, and testing datasets to quantitatively evaluate the performance of the participating algorithms.

The goal of the BraTS 2025 Glioma challenge is to identify state-of-the-art segmentation algorithms that can follow patients through the entire course of their disease. This includes aiding in initial diagnosis and treatment planning, as well as evaluating treatment efficacy and planning additional treatments. The challenge seeks to enhance clinical decision-making by providing tools that can handle the complexities of glioma imaging across different stages of treatment. Successful algorithms could significantly impact patient outcomes by improving the accuracy of tumor monitoring and enabling personalized treatment strategies.

Keywords

List the primary keywords that characterize the task.

Segmentation, Brain Tumors, Pre-treatment, Post-treatment, resection cavity, Cancer, Challenge, Glioma, Glioblastoma, MICCAI, NCI, DREAM, diffuse glioma

ORGANIZATION

Organizers

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

Jeffrey Rudie, MD PhD [Lead Organizer - Contact Person] University of California San Diego

Maria Correia De Verdier, MD, PhD Uppsala University, Sweden

Spyridon Bakas, PhD Indiana Universiy

Ujjwal Baid, PhD Indiana University

Raymond Huang, MD PhD Brigham and Women's Hospital

Evan Calabrese, MD PhD Duke University

Dominic LaBella, MD Duke University Medical Center

Rachit Saluja, MS Cornell University

Louis Gagnon, MD PhD Université Laval

Mariam Aboian, MD PhD Childrens Hospital of Philadelphia

Aly Abayazeed, MD, Stanford University.

Keyvan Farahani, PhD. National Institutes of Health

Jake Albrecht, PhD Sage Bionetworks

Verena Chung Sage Bionetworks

b) Provide information on the primary contact person.

Jeffrey Rudie, MD PhD [Lead Organizer of this task]
Department of Radiology University of California, San Diego, CA, USA Email id: jeff.rudie@gmail.com

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, and continuous evaluation after the conference deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2025 data, since our intention is to solve the particular segmentation problem, but also to provide a fair comparison among the participating methods.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

Following communication with 1) Intel 2) Neosoma Inc, and 3) Cortechs.ai we have informal confirmation for the sponsorship of monetary awards for the top 3 teams. Formal confirmation can only be provided after the acceptance of the challenge.

Note that Intel has been offering monetary awards during each of BraTS 2018-2024, Neosoma for BraTS 2021-2022, and Cortechs.ai during 2023. NIH will also provide Certificates of Merit to the top 3 performing teams.

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Coordinating the BraTS challenge proceedings with the BrainLes workshop proceedings allows the BraTS participants to publish their methods in the associated LNCS 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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 with the release of 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 Googles 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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018.
https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 BraTS 2021-2024 challenges.

All participants of the challenge will be required to accept an agreement through the synapse.org website that participation in the testing phase will automatically mean that the organizers can make the submitted containerized method publicly available through our challenge webpage.

The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Monetary awards are expected by Intel, Neosoma Inc, and Cortechs.ai
Challenge Organizers, SAGE Bionetworks, and 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

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

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.

Adult glioma patients

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 de novo diffuse gliomas of the brain, both pre-treatment and post-treatment. They will have been 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.

Multi-parametric MRI scans of the brain, including pre-contrast and contrast-enhanced T1-weighted, T2-weighted and T2-weighted FLAIR images.

Context information

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

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

The information corresponds 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 shown in 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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Tumor segmentation of the different tumor sub-regions on mpMRI.
Dice, Normalized Surface Distance, Sensitivity, Precision, Specificity - per lesion evaluation

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 [1,2,3]. Since then, multiple institutions have contributed data to create the BraTS 2025 Glioma dataset and these will be listed in a BraTS arXiv paper following acceptance of the challenge. We are currently in coordination with TCIA to make the complete BraTS 2025 dataset permanently available through their portal. All the acquisition details will be included together with the data availability in TCIA, and subsequently in IDC, including Google and AWS Marketplaces, as part of their Public Datasets Programs.

[1] U. Baid, et al., The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, arXiv preprint arXiv:2107.02314

[2] S.Bakas, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv:1811.02629

[3] Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, 4:170117 (2017) DOI: 10.1038/sdata.2017.117

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.

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 MRI scans, acquired with different clinical protocols and various scanners from: UCSF, UCSD, Duke University, Indiana University, Thomas Jefferson University, Yale University, etc.

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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 is 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 (1mm^3) and aligned to the same anatomical atlas.

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

Training data: 2800 cases (70%)
Validation data: 400 cases (10%)
Testing data: 800 cases (20%)
Total: 4000

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

The total number of cases is based on availability. The data will be split in these numbers between training, validation, and testing based on a standard split (70:10:20) used in machine learning research.

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.

Reference approved from at least 2 experienced neuroradiologists, following initial annotations from other annotation volunteers.

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 annotation of these data followed a pre-defined clinically approved annotation protocol (defined by expert neuroradiologists and/or radiation oncologists), which is provided to all clinical annotators, describing in detail instructions on what the gross tumor volume segmentation should and should not include. The annotators are 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) Enhancing Tissue (ET): This delineates the hyperintense signal of the T1-Gd, after excluding the vessels. Any areas of thick or nodular enhancement are included in the ET class, though typical treatment related thin linear enhancement along and within resection cavities and along the dura is not included in the ET class.
ii) Nonenhancing tumor core (NETC): The NETC class consists of necrotic/nonenhancing tissue surrounded by ET and not otherwise clearly represented by a prior resection cavity the necrotic core (when present). The tumor core (TC) is the union of the enhancing tumor and the necrotic core described in (i) and (ii) here.
iii) Surrounding nonenhancing FLAIR hyperintensity (SNFH): This tissue typically includes edema and infiltrating tumor. Given the post-treatment nature of the scans, any T2/FLAIR signal abnormalities, including radiation-related hyperintensity, gliosis, edema, and non-enhancing tumor, are included in the SNFH label. The Whole Tumor is the union of ET, NETC and SNFH.
iv) Resection cavity (RC): The RC class consists of both recent and chronic resection cavities. Chronic resection

cavities, which are typically older than 3-6 months, were considered those with signal intensity isointense to CSF. More recent resection cavities often contained air, blood, and/or proteinaceous materials, and otherwise exhibited variable signal characteristics.

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 is assigned to a pair of annotator-approver. Annotators span across various experience levels and clinical/academic ranks, while the approvers are board-certified neuroradiologists on the organizing committee with prior annotation experience. The annotators are 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 are satisfied with the produced annotations, these are passed to an approver. The approver is then responsible for signing off on these annotations. Specifically, the approver would review the tumor annotations, in tandem with the corresponding MRI scan, and if the annotations were not of satisfactory quality they would be sent back to the annotators for further refinements. This iterative approach is 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. The test data will be annotated and approved by two sets of annotators/approvers to assess inter-rater reliability.

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

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.

Raw MRI scans are first carefully reviewed by radiologists from individual institutions. The pre-contrast and contrast-enhanced T1-weighted, T2-weighted and T2-weighted FLAIR sequences are then extracted and named according to the standard BraTS naming convention. The four sequences are converted from their original Digital Imaging and Communications in Medicine (DICOM) file format to the Neuroimaging Informatics Technology Initiative (NIfTI) file format. Following the conversion to NIfTI files, we perform brain extraction to remove any apparent non-brain tissue (e.g., neck fat, skull, eyeballs). The brain-extracted pre-contrast and contrast-enhanced T1-weighted, T2-weighted and T2-weighted FLAIR images are registered and interpolated to the same resolution (1 mm^3). As the images are acquired over several years and multiple institutions, the software/packages used to perform these steps varied.

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
Sensitivity, Lesionwise
Specificity, Lesionwise
Precision Lesionwise

The sub-regions considered for evaluation are:

ET describes the regions of active tumor as well as nodular areas of enhancement.
NETC denotes necrosis and cysts within the tumor.
SNFH typically includes edema, infiltrating tumor and post treatment changes.
RC consists of both recent and chronic resection cavities and typically contains fluid, blood, air, and/or proteinaceous materials.
Tumor core (ET plus NETC (when present)) describes what is typically resected during a surgical procedure. Whole tumor (ET plus SNFH and NETC (when present)) as it defines the whole extent of the tumor, including the tumor core, peritumoral edematous tissue and highly infiltrated area.

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 tissue describes the regions of active tumor as well as potentially active tumor in the post-treatment setting, which in clinical practice characterizes the areas to longitudinal assess and potentially re-resect. ii) the tumor core (incl. the NETC) also what is typically resected during a surgical procedure. iii) the whole tumor as it defines the whole extent of the tumor, including the SNFH.

In terms of evaluation metrics, we use:
i) the Dice Similarity Coefficient, which is commonly used in the assessment of segmentation performance, ii) the

Normalized Surface Distance (NSD), which are both computed on a lesionwise basis per BraTS 2023 iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or under segment. 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.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses [1]. 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.

[1] S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 2: BraTS-Meningioma: Pre-operative Meningioma Tumor 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.

Meningiomas are the most common primary intracranial tumor and can be associated with significant morbidity and mortality. MRI is essential for radiologists, neurosurgeons, neuro-oncologists, and radiation oncologists in diagnosing, planning treatment, and monitoring treatment over time. The 2023 BraTS Pre-operative Meningioma (BraTS-MEN) challenge established a community standard and benchmark for state-of-the-art automated intracranial meningioma segmentation models, utilizing the largest expert-annotated multilabel pre-operative meningioma mpMRI dataset to date. Competitors in the 2023 BraTS-MEN challenge developed automated segmentation models to predict three distinct meningioma sub-regions on MRI: enhancing tumor, non-enhancing tumor core, and surrounding non-enhancing T2/FLAIR hyperintensity. Models were evaluated on separate validation and held-out test datasets using standardized metrics such as the Dice similarity coefficient and normalized surface distance.

Considering the clinical impact, the 2025 BraTS-MEN challenge will differ from the previous 2023 BraTS-MEN challenge by incorporating additional multiple expert labeled annotations for each of the cases in the 2023 BraTS-MEN challenge. The addition of multiple annotations will allow for more accurate and consistent tumor segmentations to evaluate automated segmentation models. Ground truth reference annotations for all datasets are created and approved by expert neuroradiologists for every subject included in the training, validation, and testing datasets. These data will be used by participants to develop, containerize, and evaluate their algorithms on unseen validation data until July 2025, when the organizers will stop accepting new submissions and evaluate the submitted algorithms on the hidden testing data.

The goal of the 2025 BraTS-MEN challenge is to identify state-of-the-art segmentation algorithms that can follow patients through the course of their disease including in initial diagnosis and treatment planning. The challenge seeks to enhance clinical decision-making by providing tools that can handle the complexities of meningioma automated segmentation. Successful algorithms could significantly impact patient outcomes by improving the accuracy of tumor diagnosis, monitoring, and enabling personalized treatment strategies.

Keywords

List the primary keywords that characterize the task.

Meningioma, Segmentation, Brain, Tumors, MICCAI, NCI, Artificial Intelligence

ORGANIZATION

Organizers

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

Evan Calabrese, MD PhD Duke University [Lead Organizer - Contact Person]

Dominic LaBella, MD Duke University Medical Center

Benedikt Wiestler, Technical University of Munich

Jeffrey Rudie, MD PhD, University of California San Francisco

Maria Correia De Verdier, MD, PhD Uppsala University, Sweden

Spyridon Bakas, PhD Indiana University

Ujjwal Baid, PhD Indiana University

Raymond Huang, MD PhD Brigham and Women's Hospital

Rachit Saluja, MS Cornell University

Louis Gagnon, MD PhD Université Laval

Mariam Aboian, MD PhD Childrens Hospital of Philadelphia

Aly Abayazeed, MD, Stanford University.

Keyvan Farahani, PhD. National Institutes of Health

Jake Albrecht, PhD Sage Bionetworks

Verena Chung Sage Bionetworks

Devon Godfrey, PhD Duke University Medical Center

John Kirkpatrick, MD PhD Duke University Medical Center

Zachary Reitman, MD PhD Duke University Medical Center

Chunhao Wang, PhD Duke University Medical Center

b) Provide information on the primary contact person.

Calabrese, Evan (evan.calabrese@duke.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

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2025 data, since our intention is to solve the particular segmentation problem, but also to provide a fair comparison among the participating methods.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

TBD

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018.
https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 with the release of 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release

Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 or assent has been obtained from all subjects at their respective institutions or a waiver of informed consent was approved by the local institutional review board. The protocol for releasing the data was approved by the institutional review board of the data-contributing institution. Each data contributor has acquired the necessary ethics approval.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 BraTS 2021-2024 challenges.

All participants of the challenge will be required to accept an agreement through the Synapse.org website that participation in the testing phase will automatically mean that the organizers can make the submitted containerized method publicly available through our challenge webpage.

The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Only the organizers will have access to the test image and label data.
Only the organizers will have access to the validation label data.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

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

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.

Adult, preoperative meningioma patients

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 meningioma brain tumors, 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.

Multi-parametric MRI scans of the brain, including pre-contrast and contrast-enhanced T1-weighted, T2-weighted and T2-weighted FLAIR images.

Context information

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

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

The information corresponds 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 shown in 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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Tumor segmentation of the different tumor sub-regions on mpMRI.
Dice, Normalized Surface Distance, Sensitivity, Precision, Specificity - per lesion evaluation

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).

Clinical routine diagnostic MRI. Radiotherapy planning MRI.

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

The specific case inclusion methods (pathologic, clinical/radiologic, or both) and case collection methods (i.e. retrospective, prospective, consecutive) were chosen by each participating site independently, often on the basis of pre-existing curated datasets. Imaging parameters including field strength, echo/repetition time, slice resolution, and slice thickness varied considerably between and within sites. In an effort to encourage data contribution, data contributors were not required to disclose data collection methods or MRI protocol information.

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 MRI scans, acquired with different clinical protocols and various scanners from: Duke University, University of Pennsylvania, Missouri University, University of California San Francisco, Yale University,

Thomas Jefferson University

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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.

All MRI studies in the BraTS pre-operative meningioma challenge were performed in the pre-operative and pretreatment setting and were included if one or more tumors radiographically or pathologically consistent with meningioma were included within the field of view. MRI studies containing any intracranial tumor that was not radiographically or pathologically consistent with meningioma were excluded (including cases of neurofibromatosis type 2 with intracranial Schwannomas). All cases include multiparametric MRI (mpMRI) consisting of pre-contrast T1-weighted, post-contrast T1-weighted, T2-weighted, and T2- weighted Fluid Attenuated Inversion Recovery (FLAIR) series.

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

The following estimates represent the minimum amount of pre-operative mpMRI data we intend to use for the challenge; we expect to increase these numbers through additional cohorts.
Training data: 1000 cases
Validation data: 141 cases
Testing data: 283 cases

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

Based on availability. 70% training, 10% validation, 20% testing.

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.

Each case reference annotation will be created from the fusion of at least 2 complete annotation sets that were independently approved by different experienced neuroradiologists.

The annotation annotation guidelines were based on the following:

This challenge defines 3 distinct and nonoverlapping segmentation labels for mpMRI in the pre-operative setting. These include "enhancing tumor", "nonenhancing tumor core", and surrounding non-enhancing T2/FLAIR hyperintensity (SNFH). The enhancing tumor label includes all contrast enhancing meningioma, focally thickened meninges (including dural tail), as well as en-plaque meningiomas. This label approximates the region of active, viable tumor. The non-enhancing tumor core label includes all calcification, hyperostosis, necrosis, degeneration, and any other atypical non-enhancing tumor radiographic findings. This label along with the enhancing tumor label (together comprising the "tumor core") approximately corresponds to the portion of tumor related imaging abnormality that would typically be removed in a gross total resection. The SNFH label includes the entire extent of tumor related T2/FLAIR hyperintensity surrounding the tumor core. This label is distinct from the other labels in that it is composed entirely of brain parenchyma and is not expected to contain any tumor cells, but rather represents irritated, inflamed, and/or edematous brain tissue resulting from adjacent tumor. Importantly, non-tumor related T2/FLAIR signal abnormality, commonly related to chronic microvascular ischemic.

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.

In addition to the above sub-compartment descriptions, the following common errors description was provided to all of the annotators along with figures of images with incorrect/correct label pairings for each of these descriptions.

Based on subjective review of pre-segmented meningioma cases by challenge approvers, a set of commonly encountered automated segmentation errors were identified and provided to challenge annotators in an effort to improve inter-observer variability. These commonly encountered errors included: 1. A thin rim of erroneously assigned SNFH label immediately surrounding smaller meningiomas without any true associated SNFH. 2. Incomplete or absent segmentation of small convexity meningiomas composed entirely of enhancing tumor, particularly when more than 1 meningioma was included in the field of view. 3. Improper assignment or incomplete segmentation of non-enhancing tumor regions, including exophytic hyperostosis, cystic spaces, and areas of intrinsic T1 hyperintensity, which were sometimes erroneously labeled as enhancing tumor or SNFH rather than non-enhancing tumor core. 4. Inclusion of non-tumor related brain parenchymal T2/FLAIR signal abnormality, most commonly chronic microvascular ischemic white matter changes (e.g. leukoaraiosis) within the SNFH label.

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 at least 2 pairs of annotator-approver. Annotators spanned across various experience levels and clinical/academic ranks, while the approvers were the experienced board-certified neuroradiologists.

The annotators used ITKSnap for making the annotations, and also followed 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.

TBD - Each case will utilize a specified method depending on the number and expertise of the assigned annotator-approver pairs for that respective case.. The methods will potentially consist of STAPLE, iSTAPLE, majority voting, weighted voting, expectation-maximization, consensus clustering, bayesian fusion, or fuzzy logic-based fusion.

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.

Regarding the pre-operative multi-compartment mpMRI cases:
All mpMRI data underwent standardized image pre-processing steps including conversion from DICOM to Neuroimaging Informatics Technology Initiative (NIfTI) image file format; co-registration of individual image series (T1-weighted, T2-weighted, T2-FLAIR, T1Gd) to the SRI24 atlas space including uniform 1 mm3 isotropic resampling, and automated skull-stripping using a deep convolutional neural network approach. These basic image pre-processing steps are implemented in the open-source and publicly available CaPTk and Federated Tumor Segmentation (FeTS) tool [1-3]. It should be noted that meningioma can extend through the skull and/or skull-base foramina and that any extra-cranial portions of tumors were implicitly excluded by the skull-stripping process. Despite this limitation, skull-stripping was included in the pre-processing to preserve patient anonymity (by preventing face reconstruction) and to ensure consistency with other BraTS challenges.

C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
Sensitivity, Lesionwise
Specificity, Lesionwise
Precision, Lesionwise

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

Normalized Surface Distance (NSD), which are both computed on a lesionwise basis per BraTS 2023 iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or under segment. 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.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses [1]. 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.

[1] S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 3: BraTS-Meningioma-RT: Meningioma Radiotherapy 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.

The 2024 Brain Tumor Segmentation Meningioma Radiotherapy (BraTS-MEN-RT) challenge aimed to advance automated segmentation algorithms using the largest known multi-institutional dataset of radiotherapy planning brain MRIs with expert-annotated target labels for patients with intact or post-operative meningioma that underwent either conventional external beam radiotherapy or stereotactic radiosurgery. Each case included a defaced 3D post-contrast T1-weighted radiotherapy planning MRI in its native acquisition space, accompanied by a single-label "target volume" representing the gross tumor volume (GTV) and any at-risk post-operative site. Target volume annotations adhered to established radiotherapy planning protocols, ensuring consistency across cases and institutions. For pre-operative meningiomas, the target volume encompasses the entire GTV and associated nodular dural tail, while for post-operative cases, it includes at-risk resection cavity margins as determined by the treating institution. Case annotations were reviewed and approved by expert neuroradiologists and radiation oncologists.

Considering the clinical impact, the 2025 BraTS-MEN-RT challenge will differ from the previous 2024 BraTS-MEN-RT challenge by incorporating additional expert labeled annotations for each of the cases in the 2024 BraTS-MEN-RT challenge. Fusion of multiple annotations will create more accurate and consistent tumor target label ground truths to evaluate automated segmentation models. These data will be used by participants to develop, containerize, and evaluate their algorithms on unseen validation data until July 2025, when the organizers will stop accepting new submissions and evaluate the submitted algorithms on the hidden testing data.

The goal of the 2025 BraTS-MEN-RT challenge is to identify state-of-the-art segmentation algorithms for automated radiotherapy target volume label delineation. The challenge seeks to enhance clinical decision-making by providing tools that can handle the complexities of meningioma automated segmentation. Successful algorithms could significantly impact patient outcomes by automating and standardizing meningioma radiotherapy target volumes to match or even exceed the accuracy and speed by a human annotator.

Keywords

List the primary keywords that characterize the task.

Meningioma, Segmentation, Brain, Tumors, MICCAI, NCI, Artificial Intelligence, Radiation Oncology, Radiotherapy, Stereotactic Radiosurgery, External Beam Radiotherapy

ORGANIZATION

Organizers

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

Dominic LaBella, MD Duke University Medical Center [Lead Organizer - Contact Person]

Evan Calabrese, MD PhD Duke University Medical Center

Benedikt Wiestler, Technical University of Munich

Jeffrey Rudie, MD PhD, University of California San Francisco

Maria Correia De Verdier, MD, PhD Uppsala University, Sweden

Spyridon Bakas, PhD Indiana University

Ujjwal Baid, PhD Indiana University

Raymond Huang, MD PhD Brigham and Women's Hospital

Rachit Saluja, MS Cornell University

Louis Gagnon, MD PhD Université Laval

Mariam Aboian, MD PhD Childrens Hospital of Philadelphia

Aly Abayazeed, MD, Stanford University.

Keyvan Farahani, PhD. National Institutes of Health

Jake Albrecht, PhD Sage Bionetworks

Verena Chung Sage Bionetworks

Devon Godfrey, PhD Duke University Medical Center

John Kirkpatrick, MD PhD Duke University Medical Center

Zachary Reitman, MD PhD Duke University Medical Center

Chunhao Wang, PhD Duke University Medical Center

b) Provide information on the primary contact person.

Dominic LaBella, MD Duke University Medical Center dominic.labella@duke.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

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2025 data, since our intention is to solve the particular segmentation problem, but also to provide a fair comparison among the participating methods.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

TBD

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams’ members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 with the release of 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 or assent has been obtained from all subjects at their respective institutions or a waiver of informed consent was approved by the local institutional review board. The protocol for releasing the data was approved by the institutional review board of the data-contributing institution. Each data-contributor has acquired the necessary ethics approval.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 BraTS 2021-2024 challenges.

All participants of the challenge will be required to accept an agreement through the Synapse.org website that participation in the testing phase will automatically mean that the organizers can make the submitted containerized method publicly available through our challenge webpage.

The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Only the organizers will have access to the test image and label data.
Only the organizers will have access to the validation label data.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Research, CAD, Decision support, Treatment planning, Diagnosis, Assistance, Surgery, Intervention planning, Education, Training, Radiotherapy, Stereotactic radiosurgery, External beam radiotherapy

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.

Adult meningioma patients after resection / before RT

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 meningioma brain tumors, scanned with radiotherapy planning MRI acquisition protocol including contrast-enhanced T1-weighted in native resolution.

Imaging modality(ies)

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

Radiotherapy planning MRI acquisition protocol including contrast-enhanced T1-weighted in native resolution.

Context information

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

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

The information corresponds directly to the image data

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

All patients planned to undergo radiotherapy to meningioma

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 shown in MRI 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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Tumor segmentation of the GTV on contrast-enhanced T1w images.
Dice, Normalized Surface Distance, Sensitivity, Precision, Specificity - per lesion evaluation

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).

Radiotherapy planning MRI.

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

The specific case inclusion methods (pathologic, clinical/radiologic, or both) and case collection methods (i.e. retrospective, prospective, consecutive) were chosen by each participating site independently, often on the basis of pre-existing curated datasets. Imaging parameters including field strength, echo/repetition time, slice resolution, and slice thickness varied considerably between and within sites. In an effort to encourage data contribution, data contributors were not required to disclose data collection methods or MRI protocol information.

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 MRI scans, acquired with different clinical protocols and various scanners from: Duke University, Missouri University, University of California San Francisco, SUNY Upstate Medical University, University of California San Diego, University of Washington

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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.

All MRI studies in the 2024 and 2025 BraTS-MEN-RT challenge were performed in the radiotherapy planning setting for patients with meningioma.
MRI studies containing any intracranial tumor that was not radiographically or pathologically consistent with meningioma were excluded (including cases of neurofibromatosis type 2 with intracranial Schwannomas). All cases include 3D post-contrast T1-weighted series in native resolution.

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

The following estimates represent the minimum amount of radiotherapy planning MRI data we intend to use for the challenge; we expect to increase these numbers through additional cohorts.
Training data: 500 cases
Validation data: 100 cases
Testing data: 100 cases

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

Based on availability.

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.

Each case reference annotation will be created from the fusion of at least 2 complete annotations that were independently approved by different experienced neuroradiologists and/or radiation oncologists.

The annotation annotation guidelines were based on the following:

This challenge defines a single radiotherapy target label based on a single 3D T1c MRI in the radiotherapy planning setting for intact or post-operative meningioma.

If the meningioma radiotherapy course was planned in the pre-operative setting, then the target volume label will comprise of the portion of the tumor visible on the T1c brain MRI.
If the meningioma radiotherapy course was planned in the post-operative setting, then the target will comprise of the post-op resection bed and any residual ET on the T1c brain MRI.

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.

If the meningioma radiotherapy course was planned in the pre-operative setting, then the target volume label will comprise of the portion of the tumor visible on the T1c brain MRI.
If the meningioma radiotherapy course was planned in the post-operative setting, then the target will comprise of the post-op resection bed and any residual ET on the T1c brain MRI.

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 experienced board-certified neuroradiologists and/or radiation oncologists. The annotators used ITKSnap for making the annotations, and also followed 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.

TBD - Each case will utilize a specified method depending on the number and expertise of the assigned annotator-approver pairs for that respective case. The methods will potentially consist of STAPLE, iSTAPLE, majority voting, weighted voting, expectation-maximization, consensus clustering, bayesian fusion, or fuzzy logic-based fusion.

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.

All radiotherapy planning images underwent pre-processing. This included conversion from DICOM and DICOM-RT to Neuroimaging Informatics Technology Initiative (NIfTI) image file format using dcmrtstruct2nii followed by automated defacing using the Analysis of Functional Neuroimages tool
Box. All cases used native resolution.

Robert W Cox. Afni: software for analysis and visualzation of functional magnetic resonance neuroimages. Computers and Biomedical research, 29(3):162-173, 1996

Robert W Cox and James S Hyde. Software tools for analysis and visualization of fmri data. NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo , 10(4-5):171-178, 1997.

Thomas Phil, Thomas Albrecht, Skylar Gay, And Mathis Ersted Rasmussen, Sikerdebaard/dcmrtstruct2nii:dcmrtstruct2nii v5 (version v5), 2023. URL https://doi.org/10.5281/zenodo.4037864

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise

Sensitivity, Lesionwise
Specificity, Lesionwise
Precision, Lesionwise

In terms of the assessed and evaluated tumor sub-regions:
i) For the preoperative setting, the single label for BraTS-MEN-RT will comprise the portion of the tumor visible on the T1c brain MRI.
ii) For the postoperative setting, the single label will comprise the post-op resection bed and any residual ET on the T1c brain MRI.

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 single label for BraTS MEN RT will comprise of the post-op resection bed and any residual ET on the T1c brain MRI.

In terms of evaluation metrics, we use:
i) the Dice Similarity Coefficient, which is commonly used in the assessment of segmentation performance, ii) the Normalized Surface Distance (NSD), which are both computed on a lesionwise basis per BraTS 2023 iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or under segment. 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.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses [1]. 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.

[1] S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 4: BraTS-Metastasis: Segmentation of Pre- and Post-Treatment Brain Metastases

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.

Impact: Brain metastases are the most common CNS malignancy in adults, and evaluation of brain metastases in clinical practice is commonly limited to comparison to one prior imaging study due to the common presentation of multiple metastases in a single patient. Detailed analysis of multiple patient lesions on multiple serial scans is impossible in current clinical practice because of the time it requires to assess a study. Therefore, the development of automated segmentation tools for brain metastases is critical for providing precision-based patient care. In addition, accurate detection of small metastatic lesions that are smaller than 10 mm and are an average of 1-2 mm is critical for patient prognosis, and missing even a single lesion can result in the patient requiring repeat interventions, and experience delays in treatment. In addition, gross total volume of brain metastases in a patient is an important predictor of patient outcomes and is not currently available in clinical practice due to the lack of volumetric segmentation tools that can be translated. Therefore, it is critical to develop novel segmentation algorithms for small brain metastases that detect and accurately volumetrically segment all lesions in pre-treatment and post-treatment setting, which is different from the initial 2023 edition of the challenge that was solely focused on pre-treatment segmentations. In BraTS 2025 Lighthouse Challenge, we are not only expanding the pre-treatment dataset with more diverse patients and we are adding post-treatment imaging studies, including post-radiation and post-surgical cases, but we will also independently re-annotate multiple test set cases to gauge inter-rater performance and compare algorithm performance with human (inter-rater) performance.
Many of the algorithms that were developed for gliomas, such as nnUnet, demonstrate high dice scores for larger metastases, but their performance significantly drops off for small metastases. This challenge will be critical for the development of novel segmentation and detection algorithms for brain metastases that are common in clinical practice and will provide algorithms that can be readily translated into clinical practice.

Keywords

List the primary keywords that characterize the task.

Segmentation, Brain metastasis, pre-treatment, post-treatment, contrast-enhancing lesion, peritumoral edema

ORGANIZATION

Organizers

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

Mariam Aboian, M.D., Ph.D. [Lead Organizer - Contact Person] Childrens Hospital of Philadelphia
Co-lead: Aly Abayazeed, MD Neosoma
Co-lead: Philipp Lohman, PhD Research Center Juelich (FZJ), Germany

Trainee Lead: Ahmed Moawad, M.D., Nader Ashraf, M.D.
Trainee Annotation chief: Anastasia Janas, MD/PhD

Additional coorganizers Spyridon Bakas, Kiril Krantchev, Gian Marco Conte, Fatima Memon, Florian Kofler, Ujjwal Baid, Yury Velichko, Elizabeth Schrickel, Katie Link, Hongwei Li, Sanjay Aneja, Ajay Malhotra, Ryan Maresca, Ayman Nada, Philipp Vollmuth, Victor Manuel Pérez, Keyvan Farahani, Matthew W Pease, Devon Godfrey, Scott Floyd, Jeffrey Rudie, Jake Albrecht, Verena Chung

The trainee annotator group is continuously growing, currently encompassing around 150 individuals from over 15 countries. These trainees have various backgrounds, including medical students in the latter stages of their education, radiology residents, and researchers. To ensure high-quality annotations, each one undergoes thorough training before beginning the actual annotation.

b) Provide information on the primary contact person.

Mariam Aboian MD PhD [Lead Organizer of this task] Children's Hospital of Philadelphia
mariam.aboian@gmail.com

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, and continuous evaluation after the conference deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a

collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (James Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2023 and 2024 data, since our intention is to solve the particular segmentation problem, but also to provide a fair comparison among the participating methods.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

Following communication with 1) Intel and 2) Neosoma Inc, we are currently coordinating for the sponsorship of monetary awards for the top 3 teams. Formal confirmation can only be provided after the acceptance of the challenge.
Note that Intel has been offering monetary awards during each of BraTS 2018-2022, and Neosoma for BraTS 2021.
NIH will also provide Certificates of Merit to the top 3 performing teams.

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Coordinating the BraTS challenge proceedings with the BrainLes workshop proceedings allows the BraTS participants to publish their methods in the associated LNCS 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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 with the release of 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 Googles 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

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics

and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 BraTS 2021-2024 challenges.
All participants of the challenge will be required to accept an agreement through the synapse.org website that participation in the testing phase will automatically mean that the organizers can make the submitted containerized method publicly available through our challenge webpage.
The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for "a sustainable medical imaging challenge cloud infrastructure," to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Monetary awards are expected by Intel and Neosoma Inc

Spyridon Bakas, Ujjwal Baid, SAGE Bionetworks, Mariam Aboian, MD/PhD and 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
-

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

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.

Adult brain metastasis patients

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 metastasis, clinically scanned with mpMRI acquisition protocol during pre-treatment and post-treatment 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.

Multi-parametric MRI scans of the brain, including T1w, T2w, FLAIR T2, contrast-enhanced T1w images.

Context information

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

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

The information pertains 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 in patients with brain metastases before and after initiation of treatment.

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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Tumor segmentation of the different tumor sub-regions on mpMRI.
Dice, Normalized Surface Distance, Sensitivity, Precision, Specificity - per lesion evaluation

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. Since then, multiple institutions have contributed data to create the current BraTS 2024 and upcoming BraTS 2025 Metastasis dataset and these will be listed in the latest BraTS arXiv paper following acceptance of the challenge. We are currently in coordination with TCIA to make the complete BraTS 2021-2024 dataset permanently available through their portal. All the acquisition details will be included together with the data availability in TCIA, and subsequently in IDC, including Google and AWS Marketplaces, as part of their Public Datasets Programs.

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.

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 from:
Yale University School of Medicine
Northwestern University
University of Heidelberg
University of Missouri
Duke University
Washington University
Mercy Hospital
National Cancer Institute
Stanford University
New York University (NYU)
University of California San Francisco (UCSF)
Indiana University (IU)

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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 multiple timepoints. Pre-treatment and posttreatment scans are included in the datasets. 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 (1mm^3), and aligned to the same anatomical atlas.

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

The currently available dataset include:
Total data: 3008
Training data: 1285 (additional optional 1,723 cases available for training only)
Validation data: 128 cases
Test data: 257 cases
We are working with other institutions to further increase the number of cases for the BraTS 2025 challenge.

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

Based on our preliminary data in glioblastoma and brain metastasis segmentation, the plateau for training segmentations of brain metastases is reached at approximately 150 cases using the nnUnet algorithm (Merkaj et al, 2021). The provided training data is therefore sufficient to train the algorithm. We have accumulated over 2000 cases of unlabeled data and the focus of this year's challenge is to increase the number of high-quality annotations (also on an inter-rater, per-case basis) and add about 250 further cases per month.

Merkaj S, Bousabarah K, Zeevi T, Lin M, Aboian MS. PACS-based glioma segmentation and grade prediction for clinical implementation, ASFNR-ASNR AI workshop, 2021

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.

Reference approved from at least 2 experienced neuroradiologists, following annotations from over 200 annotators (students, residents, postgraduate fellows).

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-METS 2025 challenge follows the paradigm of the BraTS 2021-2022 and BraTS-METS 2023 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 with preference placed on ITK-SNAP, 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. All segmentations are checked in the final step by one final annotator.
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.
v) resection cavity delineates the resection of region within the brain in post-treatment cases

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

Each case was assigned to a pair of annotator-approver. Annotators spanned across various experience levels and clinical/academic ranks, while the approvers are experienced board-certified neuroradiologists (with >5 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 two different approvers. Approver 1 is then responsible for signing off these annotations. Specifically, approver 1 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 approver 1) . The segmentation mask from Approver 1 is passed blindly to Approver 2 for further refinement. The whole dataset is finally approved by "Final approver" to ensure consistency. The "final approver" makes the segmentation 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 2025 challenge is similar to the one evaluated and followed by the BraTS 2017-2024 challenges. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format (Cox et al, 2004), 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 (Rohlfing et al 2010)) and interpolate 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 (Bakas et al 2017) 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 difference 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: Perform brain extraction to remove any apparent non-brain 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 (Thakur et al 2020). We then manually assessed all scans to confirm 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 (Bakas et al, 2017) (https://github.com/CBICA/CaPTk) and FeTS (https://fets-ai.github.io/Front-End/) platforms.

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, Nature Scientific Data, 4:170117, 2017. DOI: 10.1038/sdata.2017.117

T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.

R.Cox, J.Ashburner, H.Breman, K.Fissell, C.Haselgrove, C.Holmes, J.Lancaster, D.Rex, S.Smith, J.Woodward, A (Sort of) new image data format standard: NIfTI-1: WE 150, Neuroimage, 22, 2004.

S.Thakur, J.Doshi, S.Pati, S.Rathore, C.Sako, M.Bilello, S.M.Ha, G.Shukla, A.Flanders, A.Kotrotsou, M.Milchenko,

S.Liem, G.S.Alexander, J.Lombardo, J.D.Palmer, P.LaMontagne, A.Nazeri, S.Talbar, U.Kulkarni, D.Marcus, R.Colen, C.Davatzikos, G.Erus, S.Bakas, Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-institutional Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training, NeuroImage, 220: 117081, 2020. DOI: 10.1016/j.neuroimage.2020.117081

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
Sensitivity, Lesionwise
Specificity, Lesionwise
Precision, Lesionwise

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 with blood-brain barrier breakdown and based on this, clinical practice characterizes the extent of resection based on removal of the contrast-enhancing region.

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 non-enhancing 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 Normalized Surface Distance (NSD), which are both computed on a lesionwise basis per BraTS 2023 iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or under segment. 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.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses (Bakas et al, 2019). 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.

S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 5: BraTS-Africa: Segmentation of Brain Glioma in Sub-Saharan Africa patient population

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.

Brain tumors are among the deadliest types of cancer. Approximately 80% of individuals with Glioblastoma (GB) die within two years of diagnosis [1]. Brain tumors in general are challenging to diagnose, hard to treat and inherently resistant to conventional therapy. Years of extensive research to improve diagnosis and treatment of GB have decreased mortality rates in the U.S by 7% over the past 30 years [2]. Although modest, these research innovations have not translated to improvements in survival for adults and children in low- and middle-income countries (LMICs), particularly in African populations where death rates in Sub-Saharan Africa (SSA) rose by approximately 25% on average while decreasing by up to 30% in the Global North [2]. Long-term survival with GB is associated with identification of appropriate pathological features on brain MRI and confirmation by histopathology. Since 2012, the BraTS Challenge has evaluated state-of-the art machine learning methods to detect, characterize, and classify brain GB. In 2023, BraTS featured African data (BraTS-Africa) as a sub-challenge, with 18 teams participating from around the globe [3].
For 2025, the BraTS-Africa Challenge provides a renewed opportunity to expand the brain MRI GB cases from Sub-Saharan Africa in global efforts to develop and evaluate computer-aided-diagnostic (CAD) methods for detection and characterization of GB in resource-limited settings, where the potential for CAD tools to transform healthcare are more likely [4]. In particular, for the 2025 challenge we will also independently re-annotate multiple test set cases to gauge inter-rater performance and compare algorithm performance with human (inter-rater) performance.

[1] M. Poon, et al., Longer-term (>= 2 years) survival in patients with glioblastoma in population-based studies pre- and post-2005: a systematic review and meta-analysis. Sci Rep. 2020 Jul 15;10(1):11622. https://doi.org/10.1038/s41598-020-68011-4
[2] WHO GBD 2016 Brain and Other CNS Cancer Collaborators. Global, regional, and national burden of brain and other CNS cancer, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016, Lancet Neurol. 2019 Apr;18(4):376-393. https://doi.org/10.1016/S1474-4422(18)30468-X
[3] M. Adewole, et al., The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa) (arXiv:2305.19369), https://doi.org/10.48550/arXiv.2305.19369
[4] U. Anazodo, et al., AI for Population and Global Health in Radiology. Radiology: Artificial Intelligence, 2022. https://doi.org/10.1148/ryai.220107

Keywords

List the primary keywords that characterize the task.

Segmentation, Glioma, Challenge, Sub-Saharan Africa, BraTS, MRI

ORGANIZATION

Organizers

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

Udunna Anazodo, Ph.D. - Lead Organizer
McGill University

Maruf Adewole, MSc
Medical Artificial Intelligence Laboratory (MAI Lab), Lagos, Nigeria

Jeffrey Rudie, MD PhD
University of California, San Diego

Spyridon Bakas PhD & Ujjwal Baid PhD
Indiana University

Farouk Dako
University of Pennsylvania

Benedikt Wiestler
TUM School of Medicine & Health

Keyvan Farahani, Ph.D.
NIH

Jake Albrecht & Verena Chung Sage Bionetworks

Clinical Evaluators and Annotation Approvers:
Jeffrey Rudie, MD PhD
University of California, San Diego

Oluyemisi Toyobo
Crestview Radiology Ltd., Nigeria.

Olubukola Omidiji
Lagos University Teaching Hospital, Nigeria

Annotation Volunteers
Yewande Gbadamosi & Afolabi Ogunleye
Lagos State University Teaching Hospital, Lagos

Nancy Ojo
Federal Medical Centre, Abeokuta

Kator Iorpagher
Benue State University, Makurdi

Gabriel Babatunde
Lagos University Teaching

Kenneth Aguh
Federal Medical Center, Umuahia

Adaobi Emegoakor
Nnamdi Azikiwe University Hospital, Nnewi

Chinasa Kalaiwo National
Hospital Abuja

Abbas M Rabiu
Aminu Kano Teaching Hospital, Kano

b) Provide information on the primary contact person.

Udunna Anazodo PhD [Lead Organizer of this task]
Montreal Neurlogical Institute, McGill University
udunna.anazodo@mcgill.ca

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, and continuous evaluation after the conference deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data. If they do so, they MUST fully describe the additional datasets and discuss the potential difference in their results after using only the BraTS 2025 data. Since our intention is to solve the particular segmentation problem, but also to provide a fair comparison among the participating methods, enhancing the training data is conditionally permitted.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

Following communication with 1) Intel and 2) Neosoma Inc, we are coordinating for the sponsorship of monetary awards for the top 3 teams. To encourage participation and submissions from Africa and LMICs, the organizers are coordinating sponsorship through grants for prize dedicated to top teams from Africa and LMIC to attend MICCAI. Formal confirmation can only be provided after the acceptance of the challenge. Note that Intel has been

offering monetary awards during each of BraTS 2018-2023, and Neosoma for BraTS 2021-2023, and the organizers have successfully secured grant funding for the BraTS-Africa 2023 (Lacuna Fund) and 2024 (RSNA R&E; Foundation) challenges.
NIH will also provide Certificates of Merit to the top 3 performing teams.

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Coordinating the BraTS challenge proceedings with the BrainLes workshop proceedings allows the BraTS participants to publish their methods in the associated LNCS 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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics

and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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).

The brain MRI for Africa BraTS challenge is specifically retrospectively collected images where patient informed consent was not feasible. However, the study has been approved by the Institution Review Board of Western University (ID: 121287), College of Medicine of the University of Lagos (ID: CMUL/HREC/04/22/1090), Lagos State University Teaching Hospital (ID: LREC/06/10/1952), Lily Hospital Benin (ID: LH/HREC-MA/0050-23) and National Hospital Abuja (ID: NHA/EC/049/2023). Ethics approvals will also be obtained from the IRBs of all data contributing centers.
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.

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

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 BraTS 2021-2024 challenges.
The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Monetary awards are expected by Intel and Neosoma Inc.

Jeff Rudie, Spyridon Bakas, Ujjwal Baid, Maruf Adewole, SAGE Bionetworks, and 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

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

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.

Adult glioma patients

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 from Africa, diagnosed with de novo diffuse gliomas of the brain, 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.

Multi-parametric MRI scans of the brain, including T1w, T2w, FLAIR T2, contrast-enhanced T1w images.

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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Tumor segmentation of the different tumor sub-regions on mpMRI.
Dice, Normalized Surface Distance, Sensitivity, Precision, Specificity - per lesion evaluation

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 BraTS 2025 cohort has been listed in the data reference published in our related manuscripts. Since then, multiple institutions have contributed data to create the current MICCAI BraTS dataset, and these are listed in the latest BraTS-Africa arxiv paper. We are currently in coordination with TCIA to make the complete BraTS-Africa 2023/2024 dataset permanently available through their portal. All the acquisition details will be included together with the data availability in TCIA, and subsequently in IDC, including Google and AWS Marketplaces, as part of their Public Datasets Programs.

U. Baid, et al., The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, arXiv preprint arXiv:2107.02314
S.Bakas, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv:1811.02629
S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, Nature Scientific Data, 4:170117
(2017) DOI: 10.1038/sdata.2017.117
Adewole, M., Rudie, J. D., Gbadamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., Suwaid, M. A., Emegoakor, A., Ojo, N., Aguh, K., Kalaiwo, C., Babatunde, G., Ogunleye, A., Gbadamosi, Y., Iorpagher, K., Calabrese, E., Aboian, M., et Int, Anazodo, U. C. (2023). The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa) (arXiv:2305.19369). arXiv.
https://doi.org/10.48550/arXiv.2305.19369

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.

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 across sub-Saharan Africa from:
Crestview Radiology, Lagos, Nigeria (1.5 T Siemens)
Lagos University Teaching Hospital, Lagos Nigeria (1.5T Toshiba/Canon)
Lagos State University Teaching Hospital, Lagos Nigeria (1.5T Philips)
The National Hospital, Abuja, Nigeria (1.5 T Toshiba)
Aminu Kano Teaching Hospital, Lagos, Nigeria (1.5 T Siemens)
Lily Hospital, Benin (1.5T GE)
Medhub Africa (multi-scanner)
Muhimbili University of Health and Allied Sciences (1.5T and 3T multi-scanner)
Korle Bu Teaching Hospital (Toshiba 1.5T MRI)

Data Contributors:
Abiodun Fatade, MBBS
Crestview Radiology, Lagos, Nigeria

Olubukola Omidiji, MBBS
Lagos University Teaching Hospital, Lagos Nigeria.

Rachel Akinola, MBBS
Lagos State University Teaching Hospital, Lagos Nigeria

Feyisayo Daji
National Hospital, Abuja, Nigeria

M.A Suwaid, MBBS
Aminu Kano Teaching Hospital, Lagos, Nigeria

Kenneth Aguh
Medhub Africa

Mayomi Onuwaje
Lily Hospitals Benin

Ugumba Kwikima, MD
Muhimbili University of Health and Allied Sciences, Tanzania

Yaw Mensah, MBBS
Korle Bu Teaching Hospital, Ghana

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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 (1mm^3), and aligned to the same anatomical atlas.

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

Training data: 60 cases
Validation data: 20 cases
Testing data: 15 cases
These numbers are expected to increase as we continuously collect and annotate more cases. Also, we will independently re-annotate test cases to gauge inter-rater performance and estimate algorithm performance with respect to human (inter-rater) performance.

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

Based on availability.

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.

Reference approved from at least 2 experienced neuroradiologists, following annotations from 10 clinical neuroradiologists (volunteers from ASNR, Association of Radiologists in Nigeria (ARIN), or other African Imaging Societies).

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-Africa 2025 challenge follows the paradigm of the BraTS 2021-2024 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 darkened 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 2025 challenge is identical with the one evaluated and followed by the BraTS 2017-2024 challenges. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format, 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) and interpolate to the same resolution as this atlas (1 mm^3). 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 shown that the 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 difference 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, 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 non-brain 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. 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 (https://github.com/CBICA/CaPTk) and FeTS (https://fets-ai.github.io/Front-End/) platforms.

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, Nature Scientific Data, 4:170117, 2017.

T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.

R.Cox, J.Ashburner, H.Breman, K.Fissell, C.Haselgrove, C.Holmes, J.Lancaster, D.Rex, S.Smith, J.Woodward, A (Sort of) new image data format standard: NIfTI-1: WE 150, Neuroimage, 22, 2004.

S.Thakur, J.Doshi, S.Pati, S.Rathore, C.Sako, M.Bilello, S.M.Ha, G.Shukla, A.Flanders, A.Kotrotsou, M.Milchenko, S.Liem, G.S.Alexander, J.Lombardo, J.D.Palmer, P.LaMontagne, A.Nazeri, S.Talbar, U.Kulkarni, D.Marcus, R.Colen, C.Davatzikos, G.Erus, S.Bakas, Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-institutional

Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training, NeuroImage, 220: 117081, 2020. DOI: 10.1016/j.neuroimage.2020.117081

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
Sensitivity, Lesionwise
Specificity, Lesionwise
Precision, Lesionwise

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 with blood-brain barrier breakdown and based on this, clinical practice characterizes the extent of resection based on removal of the contrast-enhancing region.
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

non-enhancing 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 Normalized Surface Distance (NSD), which are both computed on a lesionwise basis per BraTS 2023 iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or under segment. 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.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses (Bakas et al, 2019). 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.

S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 6: BraTS-PEDs: Multi-Consortium International Pediatric Brain Tumor 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.

Brain tumors are among the deadliest types of cancer and the BraTS Challenge [1-3] has a successful history of resource creation for the segmentation and analysis of most common and aggressive malignant primary tumor of the central nervous system in adults, namely the glioblastoma multiforme (GBM). Although rare, pediatric tumors of the brain and central nervous system are the most common cause of disease related death in children. Brain tumors in general are challenging to diagnose, hard to treat and inherently resistant to conventional therapy because of the challenges in delivering drugs to the brain. While pediatric tumors may share certain similarities with adult tumors, their imaging and clinical presentations differs. For example, GBMs and pediatric diffuse midline gliomas (DMGs) are both high grade gliomas with short overall survival of about 11-13 months on average. GBMs are found in 3 in 100,000 people, DMGs are about three times rarer. While GBMs are usually found in the frontal or/and temporal lobes at an average age of 64 years, DMGs are usually located in the pons and often diagnose between 5 and 10 years of age. Enhancing tumor region on post-gadolinium T1-weighted MRI and necrotic region are common imaging findings in GBM. But these imaging characteristics are less common or clear in DMGs. Thus, pediatric brain tumors require dedicated imaging tools that help in their characterization and facilitate their diagnosis/prognosis. In 2022, we organized the first initiative to include pediatric brain tumors, specifically DMGs in the test set of the BraTS challenge and results were promising. These findings encouraged us to organize a larger and more diverse initiative in 2023 with multi-institutional pediatric data, leading to BraTS-PEDs 2023 challenge [4]. In BraTS-PEDs challenges, we will extend the pediatric brain tumor cohort to a larger set, collected through a few consortiums, including Children's Brain Tumor Network (CBTN) [5], DIPG/DMG-registry, and across multiple institutions. The challenge participants will have access to the pediatric training and validation data at any point from the Synapse platform. These data will be used to develop, containerize, and evaluate their algorithms in unseen validation data until December when the organizers will stop accepting new submissions and evaluate the submitted algorithms in the pediatric patient population.

[1] U. Baid, et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification", arXiv preprint arXiv:2107.02314
[2] S.Bakas, et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge", arXiv preprint arXiv:1811.02629
[3] B. H. Menze, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", IEEE Transactions on Medical Imaging 34(10), 1993-2024 (2015) DOI: 10.1109/TMI.2014.2377694
[4] Kazerooni AF, ..., Linguraru MG,. "The brain tumor segmentation (BRATS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs)". ArXiv. 2023 May 26.
[5] Familiar AM, Kazerooni AF, et al., "A multi-institutional pediatric dataset of clinical radiology MRIs by the Children's Brain Tumor Network". arXiv preprint arXiv:2310.01413. 2023 Oct 2.

Keywords

List the primary keywords that characterize the task.

Brain Tumor, Pediatric, Rare Diseases, Segmentation, Challenge, Diffuse Midline Glioma, CBTN, MICCAI

ORGANIZATION

Organizers

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

Leads:
Marius George Linguraru, D.Phil., M.A., M.Sc. - [Lead Organizer]
Children's National Hospital / George Washington University

Anahita Fathi Kazerooni, PhD, MSc - [Lead Organizer]
The Children's Hospital of Philadelphia / University of Pennsylvania

Additional members of the Organizing Team:

Zhifan Jiang, Ph.D.
Children's National Hospital

Xinyang Liu, Ph.D.
Children's National Hospital

Deep Gandhi
The Children's Hospital of Philadelphia

Nastaran Khalili
The Children's Hospital of Philadelphia

Spyridon Bakas, PhD
Department of Pathology & Laboratory Medicine, Indiana University, IN, USA

Ujjwal Baid, PhD
Department of Pathology & Laboratory Medicine, Indiana University, IN, USA

Keyvan Farahani, PhD.
National Institutes of Health

Jake Albrecht, PhD
Sage Bionetworks

Verena Chung
Sage Bionetworks

Clinical Evaluators and Annotation Approvers:
===================================
Arastoo Vossough, MD
Children's Hospital of Philadelphia

Mariam Aboian, MD
Children's Hospital of Philadelphia

Jeffrey B Ware, MD
University of Pennsylvania

Ali Nabavizadeh, MD
University of Pennsylvania

Data Contributors:
===================================
Children's Brain Tumor Network (CBTN)
DIPG/DMG Registry
Boston Children's Hospital
Yale University

b) Provide information on the primary contact person.

Marius George Linguraru (mlingura@childrensnational.org)
Anahita Fathi Kazerooni (anahitaf@upenn.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

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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.

No additional data is allowed.

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.

Organizers can participate but are not eligible for awards and will not be listed in the leaderboard.

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

TBD

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.

Three top performing teams will be announced during the MICCAI annual meeting.

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).

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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 or assent has been obtained from all subjects at their respective institutions or a waiver of informed consent was approved by the local institutional review board. 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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018.
https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 BraTS-PEDs 2024 challenge.

The National Cancer Institute takes special interest in the BraTS 2024 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for "a sustainable medical imaging challenge cloud infrastructure," to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Marius George Linguraru, Anahita Fathi Kazerooni, Spyridon Bakas, Ujjwal Baid, SAGE Bionetworks, and 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
-

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

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.

Pediatric glioma patients

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 pediatric brain tumors, 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.

Multi-parametric MRI scans of the brain, including T1w, T2w, FLAIR T2, contrast-enhanced T1w images.

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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Tumor segmentation of the different tumor sub-regions on mpMRI.
Dice, Normalized Surface Distance, Sensitivity, Precision, Specificity - per lesion evaluation

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 pediatric brain tumor images collected through CBTN have been acquired on multiple scanners, including but not limited to 1.5T and 3T Siemens and GE scanners. We expect to receive data from other institutions across CBTN and non-CBTN institutes and will provide their technical specifications in the final BraTS manuscript. Furthermore, all the acquisition details will be included together with the data availability in TCIA, and subsequently in IDC, including Google and AWS Marketplaces, as part of their Public Datasets Programs.

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.

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 from:
Children's National Hospital (CBTN site)
Children's Hospital of Philadelphia (CBTN site)
Other CBTN sites
Boston Children's Hospital
Yale University
Other sites in DIPG/DMG registry

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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 (1mm^3), and aligned to the same anatomical atlas.

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

The following estimates represent the minimum amount of data we intend to use for the challenge; we expect to increase these numbers through additional cohorts.
Training data: 260 cases
Validation data: 80 cases
Testing data: 80 cases

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

Based on availability.

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.

Reference approved by at least 2 experienced neuroradiologists, following annotations from over 30 clinical neuroradiologists (volunteers from ASNR)

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 BraTS-PEDs challenge follows the paradigm of the BraTS-PEDs 2024 challenge data. The annotation of these data followed a pre-defined clinically-approved annotation protocol (defined by a consensus of experienced pediatric neuroradiologists from the Children's Hospital of Philadelphia, with the annotation method published in [9]). This 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 [9-10] is used to produce some initial annotations followed by their manual refinements.

Summary of specific instructions (also can be found in [9]):
> Enhancing Tumor: This subregion is described by areas with enhancement (brightness) on T1 post-contrast images as compared to T1 pre-contrast. In case of mild enhancement, checking the signal intensity of normal

brain structure can be helpful.
> Cystic Component: The appearance of the cystic region is hyperintense (very bright) on T2 and hypointense (dark) on T1CE. The cystic portion should be within the tumor (versus edema which is peritumoral). The brightness is comparable to CSF.
> Non-enhancing Tumor: Any abnormal signal intensity within the tumoral region that cannot be defined as enhancing or cystic. For example, the abnormal signal intensity on T1, FLAIR and T2 that is not enhancing on T1CE should be considered as non-enhancing portion.
Edema: This sub-region is defined by the abnormal hyperintense signal (very bright) on FLAIR scans. Edema is finger-like spreading that preserves underlying brain structure and surrounds the tumor.

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 >7 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 image preprocessing pipeline applied to all the data considered in the BraTS-PEDs 2025 challenge is identical with the one evaluated and followed by the BraTS 2017-2024 challenges, with the difference in applying pediatric-specific tumor subregion segmentation and automated defacing tools. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format, 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) and interpolating to the same resolution as this atlas (1 mm3).

After completion of the registration process, we will perform automated defacing based on an in-house pediatric-specific deep-learning tool, to remove some face features that may risk re-identification of the subjects. We will then manually review all scans for confirming the correct defacing, where the complete brain region is included, and all non-brain tissue is excluded.

This whole preprocessing pipeline, and its source code are available through the CaPTk (https://github.com/CBICA/CaPTk) and FeTS (https://fets-ai.github.io/Front-End/) platforms. Pediatric-specific automated defacing and tumor subregion segmentation methods will be available on https://github.com/d3b-center/peds-brain-auto-seg-public and https://github.com/d3b-center/peds-auto-defacing-public .

C.Davatzikos, et al., "Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome." Journal of medical imaging, 5.1:011018 (2018) DOI: 10.1117/1.jmi.5.1.011018
S.Pati, et al., "The cancer imaging phenomics toolkit (CaPTk): technical overview." International MICCAI Brainlesion Workshop. Springer, Cham (2019) DOI: 10.1007/978-3-030-46643-5_38
S.Pati, et al, "The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research", Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449
Fathi Kazerooni A, et al. "Automated tumor segmentation and brain tissue extraction from multiparametric MRI of pediatric brain tumors: A multi-institutional study". Neuro-Oncology Advances. 2023 Jan 1;5(1):vdad027.
[10] Vossough A, ..., Fathi Kazerooni A, "Training and Comparison of nnU-Net and DeepMedic Methods for Autosegmentation of Pediatric Brain Tumors". Under Review (American Journal of Neuroradiology)
[11] Liu X, ..., Linguraru MG. "From adult to pediatric: deep learning-based automatic segmentation of rare pediatric brain tumors". InMedical Imaging 2023: Computer-Aided Diagnosis 2023 Apr 7 (Vol. 12465, pp. 15-19). SPIE.

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
Sensitivity, Lesionwise
Specificity, Lesionwise
Precision, Lesionwise

The regions evaluated using these metrics describe the enhancing tumor (ET), the cystic component (CC), the non-enhancing tumor (NET), the tumor core (TC), the edema (ED), and the whole tumor (WT). Note that the tumor core includes the part of the tumor that is typically resected (i.e., including ET, CC, and NET), 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 cystic component appears with hyperintense signal (very bright) on T2 and hypointense signal (dark) on T1CE.
iii) the non-enhancing tumor describes any other abnormal signal intensity within the tumorous region that cannot be defined as enhancing or cystic.
iv) the tumor core (incl. the necrotic component) describes what is typically resected during a surgical procedure.
v) the peritumoral edema defined by the abnormal hyperintense signal (very bright) on FLAIR scans.
vi) 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 Normalized Surface Distance (NSD), which are both computed on a lesionwise basis per BraTS 2023 iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or under segment. 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.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses (Bakas et al, 2019). 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.

S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 7: BraTS-GoAT: Generalizability Across Tumors

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 generating a benchmark environment and dataset for delineating
adult brain gliomas. In 2023, we expanded the dataset to include 1) the original adult gliomas population, as well as 2) the under-served sub-Saharan African brain gliomas patient population, 3) brain/intracranial meningiomas, 4) brain metastasis, and 5) pediatric brain tumor patients. This allowed us to organize different challenges, each focusing on specific clinical tasks. In this challenge, the BraTS Generalizability Across Tumors (BraTS-GoAT) Challenge, we will focus on assessing the algorithmic generalizability beyond each patient population and focus across all of them. The hypothesis is that a method capable of performing well on multiple segmentation tasks will generalize well on unseen tasks. Specifically, we aim to challenge participants to create a segmentation algorithm capable of adapting and generalizing to different scenarios with little prior information and/or data on the target class(es). We aim to simulate the clinical scenario where we develop a segmentation tool agnostic to future clinical applications (i.e., a tool trained on a specific disease(s) that will be applied to new ones without access to additional training data).

Keywords

List the primary keywords that characterize the task.

Generalizability, Segmentation, Brain, Tumors.

ORGANIZATION

Organizers

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

Baid Ujjwal, Indiana University
Conte Gian Marco, Department of Radiology, Mayo Clinic, Rochester, Minnesota, USA

b) Provide information on the primary contact person.

Conte Gian Marco (conte.gianmarco@mayo.edu)
Baid Ujjwal (ubaid@iu.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

Repeated event with annual fixed conference submission deadline.

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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.

No additional data is allowed.

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.

Organizers can participate but are not eligible for awards and will not be listed in the leaderboard.

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

TBD

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 during the annual event.

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).

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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI

Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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).

BraTS-GoAT challenge is using datasets from other BraTS challenges spanning across various tumor entities. Each data-contributor has acquired the necessary ethics approval.

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)

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), we will follow their data usage agreement.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check code availability of these challenges.

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 BraTS-GoAT 2024 challenge.

The National Cancer Institute takes special interest in the BraTS challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for "a sustainable medical imaging challenge cloud infrastructure," to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Only the organizers will have access to the test data. Sponsoring/funding: TBD

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, Intervention planning.

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.

Brain tumors patients.

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 pediatric brain tumors, 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).

The information pertains 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 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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Evaluate brain tumor segmentation algorithms' generalizability.
Dice, Normalized Surface Distance, Sensitivity, Precision, Specificity - per lesion evaluation

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).

Clinical routine MR scans.

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

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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).

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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

>16000 mpMRI from >4000 subjects will be used in this challenge.

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

Based on availability.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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.

Since the BraTS-GoAT challenge is using datasets from other BraTS challenges (specifically, Tasks 1, 2, 4, 5, and 6), please check details in the corresponding section of these individual challenges.

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.

Inter-raters variability. No difference is expected between training/validation/test sets.

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
Sensitivity, Lesionwise
Specificity, Lesionwise
Precision, Lesionwise

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

In terms of evaluation metrics, we use:
i) the Dice Similarity Coefficient, which is commonly used in the assessment of segmentation performance, ii) the Normalized Surface Distance (NSD), which are both computed on a lesionwise basis per BraTS 2023 iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or under segment. 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.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge, and 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses (Bakas et al, 2019). 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.

S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 8: BraTS-Augment: Evaluation of Augmentation Techniques for BraTS

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.

In the broader machine learning community, the concept of Data Centric machine learning has emerged to improve the performance of models with more meaningful training data. Data augmentation has been shown to improve the robustness of machine learning models, but the types of augmentations that may be useful for biomedical imaging are unknown. Conventional challenges ask participants to submit a model for evaluation on test data. This data-centric challenge will invert the process, asking participants to submit a method to augment training data such that a baseline model will show improved robustness on new data. Participants will submit a container that will augment training data (while keeping the number of training cases fixed) from the RSNA-ASNR-MICCAI BraTS 2021 (which represents the BraTS 2023 GLIOMA) challenge such that a common baseline U-Net model architecture can be trained on the container output. The trained model will be evaluated on the BraTS 2023 GLIOMA test data for Dice coefficient and normalized surface distance measures of accuracy, per lesion, with emphasis on the consistency across the test set cases. Top performing methods may offer insight to augmentation approaches that could be used to generate robust state-of-the-art segmentation models.
This challenge task will be promoted by Sage Bionetworks and PrecisionFDA, in consultation with the NCI/NIH, and the FDA Center for Devices and Radiological Health.

Keywords

List the primary keywords that characterize the task.

Segmentation, Augmentation, BraTS, Data Centric

ORGANIZATION

Organizers

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

Jake Albrecht [Lead Organizer - Contact Person] Affiliation: Sage Bionetworks

Elaine Johanson precisionFDA

Spyridon BakasIndiana University

Zeke Meier
Booz Allen Hamilton

Weijie Chen

Center for Devices and Radiological Health, U.S. Food and Drug Administration

Nicholas Petrick
Center for Devices and Radiological Health, U.S. Food and Drug Administration

Berkman Sahiner
Center for Devices and Radiological Health, U.S. Food and Drug Administration

Keyvan Farahani
National Institutes of Health

Ujjwal Baid
Indiana University

Rong Chai
Sage Bionetworks

Verena Chung Sage Bionetworks

Clinical Evaluators, Annotation Approvers, & Annotation Volunteers:
The same 65 people who facilitated the RSNA-ASNR-MICCAI BraTS 2021 challenge.

Data Contributors
The RSNA-ASNR-MICCAI BraTS 2021 challenge data contributors

b) Provide information on the primary contact person.

Jake Albrecht PhD [Lead Organizer of this task] Sage Bionetworks
jake.albrecht@sagebionetworks.org

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, and continuous evaluation after the conference deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 from publicly available datasets or 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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

Following communication with 1) Intel and 2) Neosoma Inc, we are coordinating with them for the sponsorship of monetary awards for the top 3 teams. Formal confirmation can only be provided after the acceptance of the

challenge.
Note that Intel has been offering monetary awards during each of BraTS 2018-2022, and Neosoma for BraTS 2021. NIH/NCI will also provide Certificates of Merit to the top 3 performing teams.

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Coordinating the BraTS challenge proceedings with the BrainLes workshop proceedings allows the BraTS participants to publish their methods in the associated LNCS 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.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Tool [3] (https://fets-ai.github.io/Front-End/) that offer the implementation of the evaluation metrics, as well as via the online evaluation platform (Synapse).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of Medical Imaging, 5.1:011018, 2018.

https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 have released the data in The Cancer Imaging Archive (TCIA) and the Imaging Data Commons (IDC) of the National Institutes of Health (NIH), following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions). 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 Googles 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

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018.
https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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 BraTS 2021-2024 challenges.
The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for "a sustainable medical imaging challenge cloud infrastructure," to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Monetary awards are expected by Intel and Neosoma Inc

Spyridon Bakas, Ujjwal Baid, SAGE Bionetworks, and 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

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

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.

Brain tumors patients.

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 de novo diffuse gliomas of the brain, clinically scanned with mpMRI acquisition protocol including i) pre-contrast and ii) contrast-enhanced T1-weighted, iii) T2weighted and iv) T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Imaging modality(ies)

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

Multi-parametric MRI scans of the brain, including T1w, T2w, FLAIR T2, contrast-enhanced T1w images.

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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Dice, Normalized Surface Distance, Gini index
Additional points: Sensitivity, Precision, Specificity.

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. 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. We are currently in coordination with TCIA to make the complete BraTS 2021-2024 dataset permanently available through their portal. All the acquisition details will be included together with the data availability in TCIA, and subsequently in IDC, including Google and AWS Marketplaces, as part of their Public Datasets Programs.

U. Baid, et al., The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, arXiv preprint arXiv:2107.02314

S.Bakas, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv:1811.02629

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, Nature Scientific Data, 4:170117 (2017) DOI: 10.1038/sdata.2017.117

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.
The acquisition protocol is different for each different institution as these scans we use are representative of real clinical protocols. The exact scanners and their technical specifications used for acquiring the TCIA cohort has been listed in the data reference published in the related manuscripts of ours. Since then multiple institutions have contributed data to create the current BraTS dataset and these are listed in the latest BraTS arxiv paper. We are currently in coordination with TCIA to make the complete BraTS dataset permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

U. Baid, et al., The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, arXiv preprint arXiv:2107.02314

S.Bakas, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv:1811.02629

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, Nature Scientific Data, 4:170117 (2017) DOI: 10.1038/sdata.2017.117

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 from:
1. University of Pennsylvania (PA, USA),
2. University of Alabama at Birmingham (AL, USA),
3. Heidelberg University (Germany),
4. University of Bern (Switzerland),
5. University of Debrecen (Hungary),
6. Henry Ford Hospital (MI, USA),
7. University of California (CA, USA),
8. MD Anderson Cancer Center (TX, USA),
9. Emory University (GA, USA),
10. Mayo Clinic (MN, USA),
11. Thomas Jefferson University (PA, USA),
12. Duke University School of Medicine (NC, USA),
13. Saint Joseph Hospital and Medical Center (AZ, USA),
14. Case Western Reserve University (OH, USA),
15. University of North Carolina (NC, USA),
16. Fondazione IRCCS Instituto Neuroligico C. Besta, (Italy),
17. Ivy Glioblastoma Atlas Project,
18. MD Anderson Cancer Center (TX, USA),
19. Washington University in St. Louis (MO, USA),
20. Tata Memorial Center (India),
21. University of Pittsburg Medical Center (PA, USA),
22. University of California San Francisco (CA, USA),
23. Unity Health,
24. University Hospital of Zurich.
Note that data from institutions 6-17 are provided through The Cancer Imaging Archive (TCIA - http://www.cancerimagingarchive.net/), supported by the Cancer Imaging Program (CIP) of the National Cancer Institute (NCI) of the National Institutes of Health (NIH).

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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 (1mm^3), and aligned to the same anatomical atlas.

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

Training data: 1251 cases
Validation data: 219 cases
Testing data: 570 cases

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

Based on availability.

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.

Reference approved from at least 2 experienced neuroradiologists, following annotations from 60 clinical neuroradiologists (volunteers from ASNR)

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 2025 challenge follows the paradigm of the BraTS 2021-2024 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 2025 challenge is identical with the one evaluated and followed by the BraTS 2017-2024 challenges. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format, 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) and interpolating to the same resolution as this atlas (1 mm^3). 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 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 scanners 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 difference 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, 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. 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 (https://github.com/CBICA/CaPTk) and FeTS (https://fets-ai.github.io/Front-End/) platforms.

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, Nature Scientific Data, 4:170117, 2017. DOI: 10.1038/sdata.2017.117

T. Rohlfing, et al. The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 31(5):798-819, 2010.

R.Cox, J.Ashburner, H.Breman, K.Fissell, C.Haselgrove, C.Holmes, J.Lancaster, D.Rex, S.Smith, J.Woodward, A (Sort of) new image data format standard: NIfTI-1: WE 150, Neuroimage, 22, 2004.

S.Thakur, J.Doshi, S.Pati, S.Rathore, C.Sako, M.Bilello, S.M.Ha, G.Shukla, A.Flanders, A.Kotrotsou, M.Milchenko, S.Liem, G.S.Alexander, J.Lombardo, J.D.Palmer, P.LaMontagne, A.Nazeri, S.Talbar, U.Kulkarni, D.Marcus, R.Colen, C.Davatzikos, G.Erus, S.Bakas, Brain Extraction on MRI Scans in Presence of Diffuse Glioma: Multi-institutional Performance Evaluation of Deep Learning Methods and Robust Modality-Agnostic Training, NeuroImage, 220: 117081, 2020. DOI: 10.1016/j.neuroimage.2020.117081

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
Sensitivity, Lesionwise
Specificity, Lesionwise
Precision, Lesionwise

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 Normalized Surface Distance as opposed to standard HD, in order to avoid outliers havings 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).

v) Gini index to measure case-wise distribution for DSC and NSD

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC and the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses (Bakas et al, 2019). 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.

S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 9: BraTS-Synthesis: MR Image Synthesis for BraTS

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.

Manual segmentation of brain tumors from MR images is labor-intensive and subject to significant inter-rater variability. To mitigate these issues, numerous studies have developed automated segmentation tools using deep learning (DL). These tools generally require four MRI modalities during the inference phase - namely T1-weighted (T1w) images, both with and without contrast enhancement, T2-weighted (T2w) images, and FLAIR images. However, missing MR sequences are a frequent complication in clinical practice, often due to time restrictions or image artifacts such as patient movement. Specifically, FLAIR and T1w sequences are commonly absent from routine scans. Therefore, developing methods to substitute these absent modalities is crucial for the broader adoption of these DL algorithms in clinical settings.

This task, continuing from the initial efforts in BraTS 2023 and 2024, seeks algorithms capable of substituting entire MRI volumes with generalizability to different populations and pathologies. This enables the straightforward use of BraTS segmentation networks in facilities with limited imaging protocols or for analyzing archival tumor study datasets. The challenge of reconstructing missing MRI sequences has gained increasing attention, with generative adversarial networks (GANs) showing particular promise. These algorithms must address several technical hurdles: First, there is often a discrepancy in image resolution among sequences; for instance, FLAIR images typically have anisotropic resolution when compared to the isotropic resolution of other 3D sequences. Second, some sequences may exhibit motion artifacts, while varying MRI bias fields can affect different modalities unevenly, resulting in spatially variable artifacts. Third, almost all extensive, multi-institutional datasets exhibit some degree of domain shift due to differences in acquisition parameters, scanner types, and populations (e.g. Americans, Africans, etc.). These factors must be considered when developing methods to synthesize volumetric MRI. It remains to be determined how these challenges should be managed, for example, by selecting suitable metrics or establishing invariance in the algorithms and network architecture.

In past BraTS challenges, we have introduced publicly accessible datasets and algorithms for multimodal brain glioma segmentation. In this year's MRI synthesis task, we aim to build on these foundations and previously gathered datasets to advance the development of vital computational tools for data integration and normalization. This initiative will facilitate the wider application of the tumor segmentation algorithms devised in earlier BraTS editions, applied to different pathologies (glioma, metastasis) and populations, relying on a consistent set of image modalities. The synthesis of MRI is vital for creating effective, generalizable, and reproducible methods to analyze high-resolution MRI scans of brain tumors. It will incorporate data from multiple sites, well-established in previous BraTS challenges, and introduce new inference tasks that extend beyond glioma and metastasis data. Moreover, compared to BraTS 2023/2024, at BraTS 2025, we will additionally evaluate the containerized algorithms on the underserved sub-Saharan African brain glioma patient population (BraTS-Africa) and brain/intracranial meningioma to assess their broader applicability.

Keywords

List the primary keywords that characterize the task.

Synthesis, Segmentation, Brain Tumor, Domain Generalizability

ORGANIZATION

Organizers

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

Hongwei Bran Li, [Lead Organizer]
Harvard Medical School

Benedikt Wiestler,
Technical University of Munich

Juan Eugenio Iglesias,
Harvard Medical School

Syed Muhammad Anwar,
George Washington University

Marius George Linguraru,
Children's National Hospital

Bjoern Menze,
University of Zurich

Florian Kofler,
Helmholtz Research Center

Spyridon Bakas,
Indiana University

Ujjwal Baid,
Indiana University

Jake Albrecht,
Sage Bionetworks

Keyvan Farahani,
NIH

Verena Chung,
Sage Bionetworks

Gian Marco Conte
Mayo Clinic

Clinical Evaluators, Annotation Approvers, & AnnotationVolunteers:
The same 65 people that facilitated the RSNA-ASNR-MICCAI BraTS 2021 challenge.

Data Contributors
The RSNA-ASNR-MICCAI BraTS 2021 challenge data contributors

b) Provide information on the primary contact person.

Hongwei Bran Li PhD [Lead Organizer of this task]
Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School, USA Email: holi2@mgh.harvard.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 conference submission deadline, and continuous evaluation after the conference deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other

infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2024 data, since our intention is to solve the particular segmentation problem, but also to provide a fair comparison among the participating methods.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

Following communication with 1) Intel and 2) Neosoma Inc, we are coordinating for the sponsorship of monetary awards for the top 3 teams. Formal confirmation can only be provided after the acceptance of the challenge.
Note that Intel has been offering monetary awards during each of BraTS 2018-2023, and Neosoma for BraTS 2021.
NIH will also provide Certificates of Merit to the top 3 performing teams.

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

The configuration of coordinating the BraTS challenge proceedings with the BrainLes workshop proceedings provides the BraTS participants with the option to publish their methods in the associated LNCS 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.

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 April 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 have uploaded the training and validation data to The Cancer Imaging Archive (TCIA) and the Imaging Data Commons (IDC) of the National Institutes of Health (NIH)following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions).
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 Googles AI resources. IDC data are also expected to be available through the AWS (Amazon Web Services) Marketplace.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018.
https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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

The participants must submit their containerized algorithm during or after the validation phase. Specific instructions for containerization will be provided after the challenge approval. These instructions will be very similar to what we requested participants to provide during the BraTS 2023 and 2024 challenges. The organizers will keep the containers and use them in follow-up research related to the BRATS challenge, for example, when applying new testing data available in the later BRATS challenge to enable a direct comparison of performances across the different annual editions of the BRATS challenge.

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.

Monetary awards are expected by Intel and Neosoma Inc

Spyridon Bakas, Ujjwal Baid, SAGE Bionetworks, and the organization team 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

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

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration

- Retrieval
- Segmentation
- Tracking

Synthesis, 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.

Brain tumors patients.

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 de novo diffuse gliomas of the brain, 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.

Multi-parametric MRI scans of the brain, including T1w, T2w, FLAIR T2, contrast-enhanced T1w images.

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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Participants have to generate a full image volume that corresponds to the one missing image modality (e.g., it will be one of T1w / T2w / T1c / FLAIR). Results will be evaluated regarding the accuracy of the downstream brain tumor image segmentation using Dice scores and normalized surface distance, per lesion. We will implement a BraTS algorithm (the UNet pre-trained in the FETS brain tumor segmentation initiative). The same algorithm will be used to evaluate the hidden test data. Segmentation rankings and image similarity rankings will be combined using statistical methods similar to the metric fusion approaches of previous BraTS.

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. 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 papers.

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.

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 from the segmentation task of BraTS 2022.
University of Pennsylvania (PA, USA),
University of Alabama at Birmingham (AL, USA),
Heidelberg University (Germany),
University of Bern (Switzerland),
University of Debrecen (Hungary),

Henry Ford Hospital (MI, USA),
University of California (CA, USA),
MD Anderson Cancer Center (TX, USA),
Emory University (GA, USA),
Mayo Clinic (MN, USA),
Thomas Jefferson University (PA, USA),
Duke University School of Medicine (NC, USA),
Saint Joseph Hospital and Medical Center (AZ, USA),
Case Western Reserve University (OH, USA),
University of North Carolina (NC, USA),
Fondazione IRCCS Instituto Neuroligico C. Besta, (Italy),
Ivy Glioblastoma Atlas Project,
MD Anderson Cancer Center (TX, USA),
Washington University in St. Louis (MO, USA),
Tata Memorial Center (India),
University of Pittsburg Medical Center (PA, USA),
University of California San Francisco (CA, USA),
Unity Health,
University Hospital of Zurich.

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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 (1mm^3), and aligned to the same anatomical atlas.

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

Here we will focus using the RSNA-ASNR-MICCAI BraTS 2021 dataset and the test set from BraTS-METS, BraTS-meninglioma 2024, and BraTS-Africa 2024:
Training data: 2536 cases, including 1251 cases from RSNA-ASNR-MICCAI BraTS 2021, 1285 cases from BraTS-METS.
Validation data: 347 cases, including 219 cases from RSNA-ASNR-MICCAI BraTS 2021, 128 cases from BraTS-METS.
Testing data: 992 cases, including 570 cases from RSNA-ASNR-MICCAI BraTS 2021 + 257 cases from BraTS-METS + 15 cases from BraTS-Africa + 150 meningioma

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

Based on availability.

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.

All four MRI sequences and the segmentation map will be available in the training data. In the validation and test sets, one modality out of four sequences in each case will be randomly dropped to evaluate the performance of submitted image synthesis methods.

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.

Reference approved from at least 2 experienced neuroradiologists/radiation oncologists, following annotations from 60 clinical neuroradiologists (volunteers from the BraTS 2023 annotators pool)

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.

Key annotation of all BRATS image data sets is the tumor annotation for this task.
The tumor image annotation follows the paradigm of the BraTS 2021 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:
the enhancing tumor (when present) delineates the hyperintense signal of the T1-Gd, after excluding the vessels.
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.
the tumor core, which is the union of the enhancing tumor and the necrotic core described in (i) and (ii) above.
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 2025 challenge is identical with the one evaluated and followed by the BraTS 2017-2024 challenges. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format, 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) and interpolate to the same resolution as this atlas (1 mm^3). 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 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 scanners 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 difference

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, 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 non-brain 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. 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 (https://github.com/CBICA/CaPTk) and FeTS (https://fets-ai.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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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), Lesionwise
Normalized Surface Distance (NSD), Lesionwise
SSIM, Imagewise

For glioma test set, the automated segmentation will be performed by an ensemble of SOTA glioma segmentation algorithms, including from previous BraTS challenges and FeTS. For the test set from BraTS-Mets and BraTS-Meningioma, we will use ensembles of its winner solutions. The regions evaluated with the two segmentation 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).
The Structural similarity Index (SSIM) is used to evaluate the quality of brain structures in synthetic images, i.e. to compare synthetic sequences with their physically acquired counterparts.

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

In terms of the assessed and evaluated three 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 Normalized Surface Distance as a complementary metric of overlap-based metric.
iii) the structural similarity index, which is a common perceptual metric to quantify image similarity between synthetic images and reference images.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

We will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. 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. All teams will then be placed in a ranked order and their average rankings will be randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values will be computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches. These p-values will be reported in an upper triangular matrix revealing the statistical insignificance of potential teams that will be grouped together in tiers and the significant superiority among others that we will clearly indicate. This is an evolved version of the systematic ranking that has been used on previous years for BraTS and other challenges, and will be packaged & distributed as an independent tool allowing reproducibility and use in other challenges.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for the DSC, the NSD).

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics at Indiana University School of Medicine), 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 2017-2024, uncertainties in rankings will be assessed using permutational analyses (Bakas et al, 2019). 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.

S. Bakas et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, arXiv:1811.02629 [cs, stat], Apr. 2019, Accessed: Dec. 10, 2020. [Online]. Available: http://arxiv.org/abs/1811.02629.

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

TASK 10: BraTS-Inpainting: MR Image Inpainting for BraTS

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 challenge's task is to inpaint healthy tissue in partially broken MRI scans. Reasons for this are often technical: there may be a presence of locally isolated artifacts, incompleteness of the field of view, or corrupted/missing 2D slices. For such cases, one may want to inpaint missing information locally instead of inferring the corrupted image volume completely. Therefore, our call is for algorithms capable of inpainting corrupted image intensities within a given inpainting mask. Like in the global image synthesis challenge, this will enable the application of the downstream image processing routines. For example, brain parcellation algorithms strictly require the input of normal-appearing images used in neuro-imaging studies and in brain tumor treatment planning. From a technical standpoint, these algorithms need to overcome a multitude of challenges that also apply to the global synthesis challenge:
First, the image resolutions of the individual sequences might differ; for example, FLAIR images tend to be acquired using 2D sequences, leading to anisotropic resolution, matching the resolution of other 3D imaging sequences only poorly.
Second, motion artifacts may be presented in some of the sequences. At the same time, MRI bias fields may differ in their local impact on the different image modalities, leading to spatially inconstant artifacts. Third, a general domain shift between the training and test sets due to different acquisition settings and types of scanners can be expected to be present in almost any large and multi-institutional dataset. All these effects must be considered when developing methods for synthesizing MRI locally and globally. Questions about how to deal with these challenges, for example, by choosing adequate metrics or invariance properties of the algorithms and network architecture, have yet to be answered. In previous BraTS challenges, we have set up publicly available datasets and algorithms for multi-modal brain glioma segmentation. In our challenge task for MRI synthesis, we will build on these efforts, and the previously generated data sets, to further the development of much-needed computational tools for data integration and homogenization. It will enable better integration with other downstream routines used for quantitative neuro-image analysis (that only work well for brain images without perturbations from artifacts or lesions). The resulting MRI synthesis is essential to develop effective, generalizable, reproducible methods for analyzing high-resolution MRI of brain tumors. It will include data from multiple sites well established in previous BraTS challenges, adding new inference tasks beyond image segmentation. The resulting algorithms will have the potential to benefit automated brain (tumor) image processing and improve clinical risk stratification tools for early interventions, treatments, and care management decisions across hospitals and research institutions worldwide.

Keywords

List the primary keywords that characterize the task.

Inpainting, Synthesis, Infill, Segmentation, Brain Tumor

ORGANIZATION

Organizers

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

Florian Kofler [Lead Organizer]
Helmholtz Research Center

Hongwei Bran Li
Harverd Medical School

Benedikt Wiestler,
Technical University of Munich

Juan Eugenio Iglesias
Harvard Medical School

Syed Muhammad Anwar
Childrens National Hospital

Marius George Linguraru
Childrens National Hospital

Bjoern Menze
University of Zurich

Koen Van Leemput
Harvard Medical School

Marie Piraud
Helmholtz Research Center

Spyridon Bakas
Indiana University

Ujjwal Baid

Indiana University

Jake Albrecht
Sage Bionetworks

Keyvan Farahani
NIH

Verena Chung
Sage Bionetworks

Gian Marco Conte
Mayo Clinic

b) Provide information on the primary contact person.

Florian Kofler, PhD
Helmholtz Munich, Munich, Germany
florian.kofler@helmholtz-munich.de

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, and continuous evaluation after the conference deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the

RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS data, since our intention is to solve the particular inpainting problem, but also to provide a fair comparison among the participating methods.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

Following communication with 1) Intel and 2) Neosoma Inc, we are coordinating for the sponsorship of monetary awards for the top 3 teams. Formal confirmation can only be provided after the acceptance of the challenge. Note that Intel has been offering monetary awards during each of BraTS 2018-2023, and Neosoma for BraTS.
The Brain Tumor Segmentation (BraTS) Cluster of Challenges 2021-2023.
NIH will also provide Certificates of Merit to the top 3 performing teams.

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Coordinating the BraTS challenge proceedings with the BrainLes workshop proceedings allows the BraTS participants to publish their methods in the associated LNCS post-conference proceedings.

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.

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 April 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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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 have uploaded all data in The Cancer Imaging Archive (TCIA) and the Imaging Data Commons (IDC) of the National Institutes of Health (NIH), following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions). 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 Googles AI resources. IDC data are also expected to be available through the AWS (Amazon Web Services) Marketplace.

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 [1-2], https://github.com/CBICA/CaPTk), and the Federated Tumor Segmentation (FeTS) Platform [3] (https://fets-ai.github.io/Front-End/).

[1] C.Davatzikos, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal of medical imaging, 5.1:011018, 2018. https://doi.org/10.1117/1.jmi.5.1.011018

[2] S.Pati, et al. The cancer imaging phenomics toolkit (CaPTk): technical overview. International MICCAI Brainlesion Workshop. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-46643-5_38

[3] S.Pati, et al, The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys. Med. Biol. 67(20), 204002, 2022. DOI: 10.1088/1361-6560/ac9449

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

The participants must submit their containerized algorithm during or after the validation phase. Specific instructions for containerization will be provided after the challenge approval. These instructions will be very similar to what we requested participants to provide during the BraTS 2021-2024 challenges. The organizers will keep the containers and use them in follow-up research related to the BRATS challenge, for example, when applying new testing data available in the later BRATS challenge to enable a direct comparison of erformances across the different annual editions of the BRATS challenge. The National Cancer Institute takes particular interest in the BraTS 2024 challenge and is considering providing infrastructural support in a number of ways to make the submitted algorithms available to a broader public as well. Dr. Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repositories such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Monetary awards are expected by Intel and Neosoma Inc
Spyridon Bakas, Ujjwal Baid, SAGE Bionetworks, and the organization team 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

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

Task category(ies)

State the task category(ies)

Examples:

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

Synthesis, 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.

Brain tumors patients.

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 de novo diffuse gliomas of the brain, clinically scanned with mpMRI acquisition protocol including i) pre-contrast and ii) contrast-enhanced T1-weighted, iii) T2weighted and iv) T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Imaging modality(ies)

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

Multi-parametric MRI scans of the brain, including T1w, T2w, FLAIR T2, contrast-enhanced T1w images.

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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

In the local inpainting task, participants must generate image intensities of healthy-appearing voxels that are locally voided (covering lesions or a local artifacts). Outside of these voided area(s), the full information is available. Results will be evaluated in terms of structural similarity, peak signal to noise ration and root mean square error (residual) of the image synthesized for the inpainted area and the real image. As the task is to fill in healthy appearing images, the inpainting areas of the evaluation will be localized outside of the tumor. (Unlike glioma segmentation algorithms in the global synthesis task, there is no consensus on downstream brain parcellation tasks and algorithms. To this end, we will compare brain parcellation results only in the post-challenge result analysis, and it will not contribute to the ranking.) Similarity and residual intensity-based rankings will be combined using statistical methods similar to the metric fusion approaches of previous BraTS editions.

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. 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. We are currently in coordination with TCIA to make the complete BraTS 2021-2024 dataset permanently available through their portal. All the acquisition details will be included together with the data availability in TCIA, and subsequently in IDC, including Google and AWS Marketplaces, as part of their Public Datasets Programs.

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.

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 from the segmentation task of BraTS 2022.
University of Pennsylvania (PA, USA),
University of Alabama at Birmingham (AL, USA),
Heidelberg University (Germany),
University of Bern (Switzerland),
University of Debrecen (Hungary),
Henry Ford Hospital (MI, USA),
University of California (CA, USA),
MD Anderson Cancer Center (TX, USA),
Emory University (GA, USA),
Mayo Clinic (MN, USA),
Thomas Jefferson University (PA, USA),
Duke University School of Medicine (NC, USA),
Saint Joseph Hospital and Medical Center (AZ, USA),
Case Western Reserve University (OH, USA),
University of North Carolina (NC, USA),
Fondazione IRCCS Instituto Neuroligico C. Besta, (Italy),
Ivy Glioblastoma Atlas Project,
MD Anderson Cancer Center (TX, USA),
Washington University in St. Louis (MO, USA),
Tata Memorial Center (India),
University of Pittsburg Medical Center (PA, USA),
University of California San Francisco (CA, USA),
Unity Health,
University Hospital of Zurich.

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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 (1mm^3), and aligned to the same anatomical atlas.

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

Training data: 1251 cases
Validation data: 219 cases
Testing data: 570 cases

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

Based on availability.

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.

For the local inpainting task, only the T1 sequence is used, and we provide two types of masks: tumor (unhealthy) masks and healthy masks. The tumor masks specify the regions affected by the tumor that require inpainting, while the healthy masks cover regions of the healthy brain with similar size and shape. Since there is no ground truth available for the true appearance of the tumor regions, we cannot directly evaluate the quality of inpainting in these areas. The healthy masks allow us to compute evaluation metrics in regions with known structure, providing a reliable baseline for assessing the inpainting performance.

During training, both types of masks are available to enable supervised learning. This allows algorithms to learn how to inpaint both tumor-affected and healthy regions, improving their capacity to reconstruct plausible brain structures.

In the testing phase, the focus is solely on evaluating the quality of inpainting in the healthy brain areas. All image intensities inside the inpainting regions will be set to a predefined value, and only the reconstruction of healthy areas will be used to evaluate the algorithm's performance.

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 designed an algorithm that samples inpainting masks in the healthy brain area. These areas are similar in size and shape to tumor areas. The code for obtaining these is publicly available:
https://github.com/BraTS-inpainting/2023_challenge/blob/main/dataset/dataset_generation.ipynb

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 BraTS glioma segmentation labels follow the following instructions:
Key annotation of all BRATS image data sets is the tumor annotation for this task.
The tumor image annotation follows the paradigm of the BraTS 2021 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 subregion 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:
the enhancing tumor (when present) delineates the hyperintense signal of the T1-Gd, after excluding the vessels. 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.

These glioma segmentations serve as input for our algorithm (see above) and are then curated by trained experts to make sure they do not contain other pathologies.

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 2025 challenge is identical with the one evaluated and followed by the BraTS 2017-2024 challenges. Specifically, following the conversion of the raw scans from their original DICOM file format to NIfTI file format, 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) and interpolating to the same resolution as this atlas (1 mm^3). 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 shown that the use of non-parametric, non-uniform intensity normalization (i.e., N4) to correct for intensity non-uniformities caused by the inhomogeneity of the scanners 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 difference 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, 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. 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 (https://github.com/CBICA/CaPTk) and FeTS (https://fets-ai.github.io/Front-End/) platforms.

Step 6: Run the inpainting dataset generation tool, available here:
https://github.com/BraTS-inpainting/2023_challenge/blob/main/dataset/dataset_generation.ipynb

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) and is outside the scope of the BraTS 2025 Glioma challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging, 1, 26, 2022

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)

The Structural similarity Index (SSIM) is used to evaluate the quality of brain structures in synthetic images, i.e. to compare synthetic sequences with their physically acquired counterparts, as does the L2 norm distance and Peak Signal to noise ratio (PSNR).

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

We choose the three most popular metrics to evaluate image synthesis methods:
i) RMSE (root mean square error)
ii) PSNR (Peak Signal to Noise Ratio)
iii) Structural similarity index (SSIM) which is commonly perceptual metric to quantify image similarity between synthetic images and reference images.

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.

As the metrics work on different scales we want to aggregate in scale-agnostic fashion. Further, we want to assign equal weight to all cases. The ranking scheme was developed in collaboration with Dr. Annika Reinke (DKFZ).

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, we set the scores to the worst possible rank for this case.

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

To measure the performance of the contributions, we will evaluate the quality of the infilled regions. Since ground truth data is only available for the masked regions with healthy tissue, the evaluation will be restricted to these. We will use the following set of well-established metrics to quantify how realistic the synthesized image regions are compared to real ones: structural similarity index measure (SSIM), peak-signal-to-noise-ratio, and mean-square-error (MSE). For the final ranking of the MICCAI challenge, an equally weighted rank-sum is computed across all three metrics. To compute the rank within each metric, we rank the participants for each case and again compute a rank-sum. For these computations we use challengeR.

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 2017-2024, uncertainties in rankings will be assessed using permutational analyses. Therefore, we conduct bootstrapping and robustness analysis with challengeR.

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

We want to investigate how much our rankings are driven by individual cases or in other words how robust they are.

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 addition to the challenge metrics we will also evaluate perceptual metrics such as LPIPS and provide qualitative expert evaluations.

TASK 11: BraTS-Pathology: Assessing the Heterogeneous Histologic Landscape of Glioma

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.

Glioblastoma is the most common primary parenchymal tumor of the brain. Clinically, glioblastoma has a grim prognosis with unusually short duration antecedent symptoms and median survival of 12-18 months. This malignant tumor is widely infiltrative in the cerebral hemispheres and well-characterized by heterogenous molecular profiles as well as histopathologic features. A major obstacle in treating these tumors is this molecular and micro-environmental landscape heterogeneity. Correctly diagnosing these tumors and assessing their heterogeneity is crucial for choosing the precise treatment and potentially enhancing patient survival rates. In the gold-standard histopathology-based approach to tumor diagnosis, detecting various morpho-pathological features of distinct histology throughout digitized tissue sections is crucial. Such "features" include the presence of cellular tumor, geographic necrosis, pseudopalisading necrosis, areas abundant in microvascular proliferation, infiltration into the cortex, wide extension in subcortical white matter, leptomeningeal infiltration, regions dense with macrophages, and the presence of perivascular or scattered lymphocytes. With these features in mind and building upon the main aim of the BraTS Cluster of Challenges, the goal of the BraTS-Path challenge is to develop deep-learning models capable of identifying tumor sub-regions of distinct histologic profile. These models aim to assist in the diagnosis and grading of conditions in a consistent manner. In the BraTS-Path challenge dataset, we focus on glioblastoma (GBM) digitized tissue sections with representative features. A team of neuropathologists annotated the slides, by identifying these distinct regions. Subsequently, these regions were segmented into patches classified based on the presence of specific histology. This approach established a classification task aimed at accurately identifying patches with specific features. The challenge participants can obtain the labeled training data at any point from the Synapse platform. These data will be used to develop, containerize, and evaluate their algorithms in unseen validation data until July 2025, when the organizers will stop accepting new submissions and evaluate the submitted algorithms in the hidden testing data. Ground truth reference annotations for all datasets are created and approved by expert neuropathologists for every subject included in the training, validation, and testing datasets to evaluate the performance of the participating algorithms quantitatively.

Keywords

List the primary keywords that characterize the task.

Classification, Pathology, Digital Pathology, Brain Tumor, Cancer, Challenge, Glioma, Glioblastoma, health disparities, MICCAI, NCI, DREAM, diffuse glioma

ORGANIZATION

Organizers

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

Spyridon Bakas [Lead Organizer - Contact Person]
Indiana University
Jake Albrecht
Sage Bionetworks
Verena Chung
Sage Bionetworks
Lee A D Cooper
Northwestern University
Shahriar Faghani
Mayo Clinic
Keyvan Farahani
NIH
Mana Moassefi
Mayo Clinic
Sarthak Pati
Indiana University
Siddhesh Pravin Thakur
Indiana University
Clinical Organizers:
Robert Bell,
Indiana University
Jason Huse,
MD Anderson Cancer Center

b) Provide information on the primary contact person.

Spyridon Bakas, PhD
[Lead Organizer of the "BraTS-Pathology: Assessing the Heterogeneous Histologic Landscape of Glioma"]
Indiana University
Email: spbakas@iu.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 conference submission deadline, and continuous evaluation after the conference deadline

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2025 data, since our intention is to solve the particular classification problem, but importantly to provide a fair comparison among the participating methods.

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 cannot be 1) eligible for awards, 2) announced as the winners of the challenge, or 3) included in the announced formal rankings. They will however be evaluated and if they are within the top-ranked ones they will be honorarily mentioned to contribute back to the community. 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 official leaderboard.

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

The lead organizer of the challenge is in communication with Intel and AWS, to sponsor monetary awards for the top 3 teams. Formal confirmation can only be provided after the acceptance of the challenge. Note that Intel has been offering monetary awards during each of BraTS 2018-2024.
NIH will also provide Certificates of Merit to the top 3 performing teams.

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 during an oral presentation.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

The configuration of coordinating the BraTS challenge proceedings as Springer LNCS proceedings provides the BraTS participants with the option to publish their methods in 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.

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 publicly release the training set in April 2025 together with the validation 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. However, only 2 submissions will be allowed in the final testing/ranking data/phase to avoid potential tuning of the submitted approach to the testing data.

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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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).

Not applicable, as these cases are already publicly available from The Cancer Imaging Archive (TCIA), as part of the TCGA-GBM and TCGA-LGG data collections. However, please note that expert clinicians at our end worked on their reclassification according to the latest WHO classification criteria.

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, as the exact training/validation data will follow the existing TCGA-GBM and TCGA-LGG 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 Generally Nuanced Deep Learning Framework (GaNDLF - https://github.com/mlcommons/GaNDLF).

Pati S, Thakur SP, Hamamci E, Baid U, Baheti B, Bhalerao M, et al. GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows. Communications Engineering. 2023;2(1):1-17.

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 BraTS 2021-2024 challenges. All participants of the BraTS-Path challenge they will be required to accept an agreement through the synapse.org website that participation in the testing phase of the challenge, will automatically mean that we can make their

containerized method publicly available through our challenge webpage.
The National Cancer Institute takes special interest in the BraTS challenge and is considering providing infrastructural support in several ways. Dr Keyvan Farahani, a long time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC) on the Google Cloud Platform.

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.

Monetary awards are expected by Intel and AWS
Spyridon Bakas, Shahriar Faghani, Siddhesh Thakur, SAGE Bionetworks (synapse.org), and the clinical evaluators will be the only ones who will have access to the validation, and test case ground truth 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

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

Task category(ies)

State the task category(ies)

Examples:

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

Classification

Cohorts

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

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

Patients, diagnosed with de novo diffuse gliomas of the brain.

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 de novo diffuse gliomas of the brain, with clinically digitized tissue sections using the paradigm of Formalin-Fixed Paraffin-Embedded (FFPE) and stained with Hematoxylin and Eosin.

Imaging modality(ies)

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

Histopathology images. Specifically, H&E-stained; FFPE digitized tissue sections.

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 class)

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 tumor tissue showing in H&E-stained; FFPE digitized tissue sections.

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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy, AUC, F1 Score, Matthews Correlation Coefficient (MCC), Sensitivity, Specificity.

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 H&E-stained; FFPE digitized tissue sections used in the BraTS-Path challenge, describe histology images acquired during standard clinical practice across the 11 International sites mentioned in (c) below. The exact staining process details and the digital scanners (with their technical specifications) used for acquiring this TCIA cohort are not publicly available neither through TCIA, nor through the Genomic Data Commons (GDC) Data Portal of the NIH/NCI.

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

The exact staining process details and the digital scanners (with their technical specifications) used across these 11 sites to acquire this TCIA cohort are not publicly available neither through TCIA, nor through the Genomic Data Commons (GDC) Data Portal of the NIH/NCI.
We appreciate that the acquisition protocols and equipment are different across (and within each) contributing

institution, as these represent real routine clinical practice. We are in coordination with TCIA to identify as much of these specific details for each image of each patient and then publish this as supplementary material together with the challenge meta-analysis manuscript.

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 H&E-stained; FFPE digitized tissue sections, acquired with different clinical protocols and various scanners from:
1) Henry Ford Hospital (MI, USA),
2) University of California (CA, USA),
3) MD Anderson Cancer Center (TX, USA),
4) Emory University (GA, USA),
5) Mayo Clinic (MN, USA),
6) Thomas Jefferson University (PA, USA),
7) Duke University School of Medicine (NC, USA),
8) Saint Joseph Hospital and Medical Center (AZ, USA),
9) Case Western Reserve University (OH, USA),
10) University of North Carolina (NC, USA),
11) Fondazione IRCCS Instituto Neuroligico C. Besta, (Italy),
Note that data from these institutions are provided through The Cancer Imaging Archive (TCIA - http://www.cancerimagingarchive.net/), supported by the Cancer Imaging Program (CIP) of the National Cancer Institute (NCI) of the National Institutes of Health (NIH).

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).

Clinical experts (neuropathologists and technicians) involved in tissue staining for suspected and diagnosed brain tumor patients 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.

Based on the given definition (in this section - (a)) that a case encompasses data processed to produce one result that is compared to the corresponding reference result, a case in this challenge represents an individual patch extracted from an H&E-stained; FFPE digitized tissue section of a single patient tumor at a specific timepoint. We ensured that the patches were of a similar size, with each representing either a specific class present in that patch or none, in which case it was classified as 'background'. These tissue sections exhibit a variety of features indicative of the diagnosis of a glioblastoma and have been annotated by expert neuropathologists. These annotated regions are divided into same size patches, each of them corresponding to a distinct morpho-histologic feature (or class) that the participants are expected to predict. Since this task has not been conducted before, we consider individual patches as individual cases in this challenge, with the intention of conducting a deeper analysis and offer a deeper understanding of these distinct features/classes in a more fine-grained resolution. Specifically, we would like to assess the intrinsic similarity of these classes and hence inherent difficulty of detecting individual classes, as well as which are the most confused with each other features/classes. Throughout both the training, validation, and testing phases, these patches are classified according to their respective features/classes. The inclusion criteria for each tissue section were determined by the presence of histologic features characteristic of glioblastoma. Please note that all tissue section included for each case of the provided dataset, represent the tissue sections with the best quality available for this particular case.

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

Training Data: 150,000
Validation Data: 100,000
Testing Data: 200,000

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

Based on availability. All available data were split into training, validation, and testing following a 70%-10%-20% proportion, in line with conventional proportions used in machine learning studies.

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.

Class distribution in classification tasks chosen according to real-world distribution. Choice was made to ensure methods that can consider the challenging real-world problem and extend to a more difficult task in the next year.

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.

Reference approved or edited until consensus from 2 experienced neuropathologists, following manual annotations from 16 clinical neuropathologists (volunteers from the RANO cooperative group)

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 annotation of these data followed a pre-defined clinically-approved annotation protocol (defined by expert neuropathologists), which was provided to all clinical annotators, describing in detail instructions on what the segmentations of each histologic feature 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, or the provided infrastructure based on the Digital Slide Archive available through a web portal by Indiana University,

and follow a complete manual annotation approach.
Summary of specific histologic areas of interest:
i) presence of cellular tumor
ii) pseudopalisading necrosis
iii) areas abundant in microvascular proliferation
iv) geographic necrosis
v) infiltration into the cortex
vi) penetration into white matter
vii) leptomeningeal infiltration
viii) regions dense with macrophages
ix) presence of lymphocytes

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 neuropathologists (with >10 years of experience), listed in the Organizers' section as Clinical Organizers. The annotators were given the flexibility to use their tool of preference for making the annotations, or the provided infrastructure based on the Digital Slide Archive available through a web portal by Indiana University, and follow a complete manualannotation approach. 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 tissue section, and the annotations of not satisfactory quality were removed from the provided annotation. If the patches from the remaining annotations were less than approximately 1,500 patches then the tissue sections would be sent back to the annotators for further annotations. 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 cases.

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 TCGA-GBM and TCGA-LGG data sets, which are publicly accessible via the TCIA, have been chosen for this challenge. Initially, we have reclassified these collections in line with the 2021 WHO classification of CNS tumors. This reclassification was specifically done to pinpoint all cases of GBM IDH-wildtype, which are categorized under CNS WHO grade 4. The TCGA-LGG collection, initially classified as low-grade astrocytomas, is redefined under the 2021 WHO CNS criteria as GBM due to specific molecular characteristics indicative of distinct tumor evolution. Consequently, these astrocytomas, now classified as molecular GBM, are included in this challenge to develop algorithms applicable to all clinical GBM as per WHO guidelines. Conversely, certain cases in the TCGA-GBM collection have been excluded because their molecular profiles do not align with the current WHO definition of

GBM. For this study, multiple H&E-stained; tissue sections from each case in the reclassified TCGA-GBM and TCGA-LGG collections are used. Focusing solely on Formalin-Fixed Paraffin-Embedded (FFPE) slides, we avoid hydration artifacts common in frozen sections. Post annotation of histologically distinct regions by clinical experts, each region is segmented into 512x512 patches. No patch-level image curation is essential as annotations are created carefully on areas of high confidence for clear appearance content

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.

N/A, as only areas of high confidence from at least 2 neuro-pathologists are used for the patch creation. Perhaps at a later version of the challenge we could propose to study and evaluate the effect of any potential annotations error as an uncertainty task, similar to the one we did in BraTS 2019-2020 (i.e., to quantify the uncertainty in the tumor segmentations), but for now this is outside the scope of the BraTS-Path 2025 challenge.

R.Mehta, et al, QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning Biomedical Imaging, 1, 26, 2022

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)

Accuracy,
AUC,
F1 Score,
Matthews Correlation Coefficient (MCC),
Sensitivity,
Specificity.

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

In terms of evaluation metrics, we use:
I) Accuracy will provide the proportion of true results (both true positives and true negatives) among the total number of cases examined. It proves the effectiveness of the classification model.
II) MCC provides a balanced measure even when the classes are of very different sizes. It is a correlation coefficient between the observed and predicted classifications, offering a more informative and nuanced

assessment than simple accuracy.
III) As a measure that balances precision and recall (sensitivity), the F1 score is crucial for scenarios where the cost of false positives and false negatives is high. It is particularly useful when dealing with imbalance between classes.
IV) AUCROC curve, quantifies the overall ability of the model to discriminate between the positive and negative classes across different thresholds. A higher AUC indicates better model performance
V) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or underclassify different classes.

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.

Ranking of methods will be based on the aggregate of rankings of F1 score and MCC. Specifically, we will follow the DELPHI-based recommendations for image analysis validation [1,2], incorporating i) algorithmic ranking, and ii) statistical significance testing. Building upon the approach in BraTS 2017-2024, the performance of the classification task will be evaluated based on the relative performance (as an aggregate metric of the ones described above) of each team's model in classifying different tumor tissue types. For this analysis we will divide the test data into multiple non-overlapping subsets, ensuring balanced class representation. The assessment will involve a detailed analysis of each model's classification performance for various tumor tissues. We then compute an average rank for each of the subsets across both metrics, and aggregated these average rankings to produce a conclusive overall ranking.

[1] Reinke et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024 Feb;21(2):182-194.
[2] Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195-212.

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 (e.g., 0 for accuracy and F1 score)

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

Following discussions with the biostatistician involved in the design of this challenge (Dr Kun Huang, Chair of Dept of Biostatistics & Health Data Science), 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.

All approaches are placed in a ranked order using the overall ranking score described above, and their average rankings are randomly permuted (i.e., 500,000 permutations), in a pair-wise manner. Corresponding pairwise p-values are computed to determine the pair-wise statistical significance and report actual differences between the ordered ranked approaches.

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

TASK 12: BraTS-Pro: Brain Tumor Progression Challenge

SUMMARY

Abstract

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

Brain tumors are among the most researched diseases in the field of medical image computing. This is reflected by the popularity of challenges like the MICCAI BraTS [1], the MICCAI FeTS challenge [2], and the use of brain MRIs in challenges like MICCAI MOOD [3] for anomaly detection. Methods for semantic segmentation developed in this context show astonishing results, even comparable to intra- and inter-rater variability [4], with only marginal differences between the top performing participants. However, most of the challenges and further research focus only on single time-points. Longitudinal properties are usually only inferred from the single time-point segmentation results if needed (e.g. [5]). These longitudinal properties play a huge role in the field of brain tumor research, because they can be used for assessing treatment response. The RANO (Response Assessment in Neuro Oncology) working group [6] defines different types of response, namely complete response, partial response, stable disease and progressive disease. Progressive disease by contrast enhancing lesions is defined as an increase of the product of perpendicular diameters enhancing tumor lesions by at least 25% [6] or the appearance of any newly formed enhancing lesion [6]. The early detection of these kinds of progression in brain tumor patients is crucial for further treatment decisions, as well as assessing the response to drugs, e.g. in clinical studies.

The two kinds of progression can be extracted in a (semi)-automatic manner from the segmentations of the individual time points.

The gain in tumor volume can be extracted from automated segmentation on the individual time points, by translating the threshold on tumor growth to an increase of 40% in tumor volume [4,8]. These measurements are already optimized on the single scans.

Detection of newly formed lesions from automated segmentations is a significantly more sophisticated procedure. It involves registration between consecutive scans (for difficulties regarding longitudinal registration see [7]), the definition of "volume at risk" (i.e. volume where a new lesion might appear) and finally distinguishing newly formed lesions in this volume from lesion growth from existing lesions. A more in-depth description of this procedure is given in [8].

This is further complicated by the optimization targets of common semantic segmentation methods. On the one hand, they do not directly optimize for the detection of individual lesions, posing the risk that a newly formed lesion is not detected by the network. On the other hand, they often do not take longitudinal information into account, ignoring important information.

Finally, automated segmentations do not properly cover non-measurable lesions, which play an important role in the qualitative response assessment.

An end-to-end approach for the detection of disease progression, circumventing manual interventions, and tumor segmentations would therefore be preferable.

To this end, we propose the second iteration of the Brain Tumor Progression Challenge (BraTPRO) as part of the

BraTS25 cluster of challenges to address this gap in current research. The challenge directly tackles the classification of the different types of response according to RANO criteria (complete response, partial response, stable disease, progressive disease - see [6] for more details about the different types of response).

Participants will develop and train their method with all data available to them - including our provided public dataset with response classification annotations and any other accessible data sources. Their final model will be submitted, and inference will be executed on the organizers' hidden test set. As the hidden test set is proprietary, this evaluation represents a realistic scenario with an unknown distribution shift between the participants' training data and our test data.

The publicly available longitudinal dataset provided to the participants is the LUMIERE dataset [9]. It comprises 91 patients with a total of 616 scans and annotations according to RANO criteria. Our hidden test dataset is the multicentric EORTC-26101 dataset. This dataset comprises 306 patients with glioblastoma, as previously reported in [8], for which we provide response classification annotations according to RANO criteria and ground truth segmentation masks. We split this dataset into 300 cases that are reserved for the final test phase and 6 cases that are used for validation.

In the first phase of the challenge, participants are provided the LUMIERE dataset with response classification annotations and the automatically generated segmentations. During this phase participants will develop their models and can submit them for validation on our validation set.
In the final phase participants submit their trained model for a final evaluation on the 300 test cases of the EORTC-26101 dataset.

We hope that this challenge will raise awareness of the gap in current research related to longitudinal properties beyond the field of brain tumor research.

Keywords

List the primary keywords that characterize the task.

Brain Tumors, Disease Progression, RANO, Longitudinal Image Analysis

ORGANIZATION

Organizers

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

Yannick Kirchhoff, Tassilo Wald, Balint Kovacs, Maximilian Zenk, Klaus Maier-Hein
German Cancer Research Center (DKFZ), Heidelberg, Division of Medical Image Computing, Germany

Philipp Vollmuth
Division for Computational Radiology and Clinical AI, Clinic for Neuroradiology, University Hospital Bonn, Germany
Division of Medical Image Computing, German Cancer Research Center, Heidelberg, Germany

Martha Foltyn-Dumitru
Division for Computational Radiology and Clinical AI, Clinic for Neuroradiology, University Hospital Bonn, Germany

Jens Kleesiek, Jan Egger
Institute for AI in Medicine (IKIM), University Hospital Essen, Germany

Yannick Suter, Mauricio Reyes
ARTORG Center, University Bern, Switzerland

André Ferreira
Center Algoritmi, University of Minho, Braga, Portugal
Institute for AI in Medicine (IKIM), University Hospital Essen, Germany

Spyridon Bakas (Indiana University, Indianapolis, IN, USA)
Raymond Y Huang (MGB, Boston, MA, USA)
Javier Villanueva-Meyer (University of California San Francisco, San Francisco, CA, USA)
AI-RANO Group leads

b) Provide information on the primary contact person.

Philipp Vollmuth
p.vollmuth@dkfz-heidelberg.de

Yannick Kirchhoff
yannick.kirchhoff@dkfz-heidelberg.de

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.
However, we plan to enable further testing of methods on the test set with a separate post-challenge leaderboard. Additionally, we plan to repeat this challenge at future MICCAIs.

Challenge venue and platform

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

MICCAI

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

Synapse.org
Following our successful collaboration with the Synapse platform (SAGE Bionetworks) since the RSNA-ASNR-MICCAI BraTS 2021 challenge, we have coordinated with them and following the support from NCI (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 cluster of challenges. The National Cancer Institute takes special interest in the BraTS 2025 challenge and is considering providing infrastructural support in a number of ways. Dr Keyvan Farahani, a long-time co-organizer of BraTS challenges and a project scientist on a collaborative NCI Informatics Technology for Cancer Research (ITCR) grant, is the recipient of an NIH Office of Data Science and Strategy (ODSS)-STRIDES award for a sustainable medical imaging challenge cloud infrastructure, to further implement open (continuous) challenges by supporting cloud compute and other infrastructures for (a) benchmarking of tools and automated submission of containerized tools for evaluation, (b) hosting of top-ranking tools through NCI FireCloud Resource and public tool repository such as Dockstore or ModelHub, and (c) hosting resulting image annotations as derived data in the Imaging Data Commons (IDC). All aforementioned NCI platforms are implemented on the Google Cloud Platform. This collaboration with Synapse, enabled by NCI/NIH support through ITCR grant (Jamed Eddy, PI) and other NCI resources represents a major advancement in the challenge design and leveraging of public resources.

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

https://www.synapse.org/brats2025 - (Website will be publicly visible after the challenge approval)

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 allowed to use additional data from publicly available datasets and their own institutions, for further complementing the data, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2025 data, since our intention is to solve the particular classification problem, but importantly to provide a fair comparison among the participating methods.

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 not participate

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

TBD

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The top 3 performing methods will be announced publicly at the conference if they don't opt-out and the teams will be invited to present their method. All teams are free to decide if they want to show up on the public leaderboard.

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).

Teams will be invited to nominate members as co-authors if they:
- follow all the rules of the challenge
- open-source their algorithm
The organizers reserve the right to exclude teams or members from the author list in case of violation of these rules.
Submissions of teams which decide to not nominate anyone as co-authors will still be used in the publication.
We do not limit the number of co-authors per team.
The participating teams are encouraged to publish their methods separately and we will not enforce an embargo time.

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.

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.

The dataset used for training is the public Lumiere dataset, allowing participants immediate access to start developing their methods. Multiple submissions to the online evaluation platform will be allowed for the validation phase, enabling participants to validate their solution(s). However, only 2 submissions will be allowed in the final testing/ranking data/phase to avoid potential tuning of the submitted approach to the testing data.

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).

1 March 2025: Registration opens
Participants will be able to register for the challenge in synapse.org, from the date of its potential acceptance (March 1, 2025) until the short paper submission deadline (July 31, 2025).

1 April 2025: Training and validation data release
Availability of training data (with ground truth labels) and validation data (without ground truth labels).

31 July 2025: Short paper submission deadline
Reporting method & results on training and validation data. The only difference with the final paper submission should be the inclusion of the testing results in the camera-ready submission

15 August 2025: Containerized algorithm submission deadline.
Evaluation on testing data by the organizers, only for participants with submitted short papers. Ranking of all participating methods, following statistical significance assessment based on multiple permutation testing.

22 August 2025: Invitation to participate
Inviting all participants with valid submissions (paper + container) to present at the conference (type of presentation will be determined within the next 2 weeks)

1 September 2025: Contacting top-performing methods for preparing slides for oral presentation.

23-27 September 2025: Challenge at MICCAI
Announcement of final top 3 ranked teams

15 October 2025: Camera-ready submission deadline
Incl. results on testing data, for inclusion in the associated LNCS proceedings.

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).

The training set relies on the already publicly available dataset, for which the cantonal ethics committee of Bern (Switzerland) approved the studies and waived written informed consent.

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)

Public training set: CC BY-NC
Validation and test set are not published

Code availability

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

The evaluation code for the challenge will be made publicly available on GitHub.

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

Participating teams have to open-source their code under a license which allows public use (preferably CC-BY license) in order to be eligible to win the challenge.

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.

The challenge may obtain dedicated sponsoring or funding, which will not have influence on challenge design, evaluation and results.
Only members of the organizers' departments have access to the test labels.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

- Diagnosis

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

Decision support, Longitudinal study, Research, Diagnosis

Task category(ies)

State the task category(ies)

Examples:

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

Classification

Cohorts

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

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

Patients with a previous medical record of glioma and multiple consecutive MRI scans, pre- and/or posttreatment. At least one follow-up examination.

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

Patients with a previous medical record of glioma and multiple consecutive MRI scans, pre- and/or posttreatment. At least one follow-up examination.

Imaging modality(ies)

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

Multi-parametric MRI scans of the brain, including T1w, T2w, FLAIR T2, contrast-enhanced T1w images.

Context information

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

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

Training: The training dataset contains many additional information in addition to the MRI scans like MR acquisition parameters and time between scans. Notably automatically generated segmentations using two tools are given for each scan.
Validation and Testing: None

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

Training: The training set contains metadata on the patients regarding, for example, patient sex, overall survival, and age.
Validation and Testing: 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 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.

Patients with brain tumors, scanned with clinically routine MRI.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.

- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Precise classification of response according to RANO.

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).

Training data: The training data, the LUMIERE dataset [9], stems from the Bern University Hospital (Inselspital), the pre-operative scans were acquired between 2008 and 2013, follow-ups were recorded until 2017.
95% of the 2487 provided MRI images have been acquired on Siemens scanners, 3% on Philips scanners (Philips Medical Systems/Philips Healthcare), and 2% on scanners from GE Medical Systems. Information on the respective scanner is available for all scans.
Validation and Testing data: n/a

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

Training data: Image acquisition parameters are given for all individual scans.
Validation and Testing data: n/a

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.

Training data: The development data was acquired at the Bern University Hospital (Inselspital).
Validation and Testing data: Multi-center dataset, more precise information on included centers can be shared with reviewers if they agree not to participate in the challenge.

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).

Clinical staff involved in MRI acquisition for suspected and diagnosed brain tumor patients 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.

Cases consist of two consecutive brain MRI scans of a single patient. The provided sequences are unenhanced and contrast-enhanced T1-weighted, T2-weighted and T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Training data: 91 patients with a total of 616 scans. Some scans do not contain all sequences.
Validation data: 6 patients with 2 scans each resulting in a total of 12 scans
Testing data: 300 patients with 2 scans each resulting in a total of 600 scans

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

The provided training data is the full publicly available dataset.
The validation and testing data is part of a private dataset. The validation set is important to gauge the performance of trained models under the distribution shift but is limited in size to avoid overfitting on the validation set.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

There is a (possible) distribution shift between the training dataset on the one hand (data from a single center) and testing cases on the other hand (data from several centers and scanners). In addition to the acquisition shift there might also be a shift in the class distributions as well as a population shift. Information on these will however not be made available for participants.

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.

Training dataset: Manual image annotation by one annotator
Testing dataset: Manual image annotation by two annotators

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.

Training dataset: No specific instructions beyond the RANO criteria.
Testing dataset: No specific instructions beyond the RANO criteria.

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.

Training dataset: expert neuroradiologist with 14 years experience
Testing dataset: n/a

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.

Training dataset: n/a
Testing dataset: Consensus discussion

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.

Training data: Scans are converted to Nifti file format and skull-stripped using the HD-BET [13] tool.
Testing data: Scans are converted to Nifti file format, skull-stripped using the HD-BET [13] tool and registered to the T1 volumes. Scans are not resampled to a common spacing in order to keep it close to the clinical workflow.

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.

The annotations are done by experienced radiologists. However, even though the RANO categorisation aims to make assessment objective, there is still a potential source of error related to ambiguous cases.

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)

Balanced Accuracy (BA), F1-score, TPR, TNR, AP, AUROC

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

The metrics were chosen following the guidelines from the Metrics Reloaded Framework [14] for an imbalanced dataset (BA, F1-score, AP) with further metrics added for additional insights (TPR, TNR, AUROC).

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.

Submissions will be ranked using F1-score, BA and AP for all classes separately. The final ranking is based on the respective ranks by averaging over classes and metrics, with weights of 0.25/0.25/0.5 for F1-score, BA and AP, respectively.

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

As participants submit docker images for testing there should not occur any missing results. If an algorithm should still fail to produce a result for a specific case, we will assign the worst possible result to this case, i.e. a wrong label with lowest score for the true class.

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

AP, BA and F1-score represent multi-threshold and counting metrics, respectively. The used ranking scheme will ensure that winning methods need to perform well on a fixed cutoff as well as on moving thresholds and outperform other methods on all classes.

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.

Variability of rankings will be assessed by bootstrapping methods.

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

Bootstrapping is among the suitable methods for the assessment of ranking variability according to [15].

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.

We plan to further analyze patterns in the predictions to find hard cases in the dataset and investigate the effect of the distribution shift between training and testing data.
In addition, we will investigate ranking variability using bootstrapping.

References:

1] Bakas, S. et al. "The International Brain Tumor Segmentation (BraTS) Cluster of Challenges"
doi: 10.5281/zenodo.7837973
[2] Bakas, S. et al. "The Federated Tumor Segmentation (FeTS) Challenge 2022"
doi: 10.5281/zenodo.6362408
[3] Zimmerer, D. et al. "Medical Out-of-Distribution Analysis Challenge 2023"
doi: 10.5281/zenodo.7845019
[4] Menze, B. et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)" IEEE Transactions On

Medical Imaging 34, 1993-2024 (2015)
doi: 10.1109/TMI.2014.2377694
[5] Menze, B. et al. "Proceedings of MICCAI-BRATS 2016"
https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2016_proceedings.pdf
[6] Wen PY, Macdonald DR, Reardon DA, et al. "Updated response assessment criteria for high-grade gliomas: Response Assessment in Neuro-Oncology Working Group." Journal of Clinical Oncology 28:1963-1972 (2010)
[7] Baheti, B. et al. "The Brain Tumor Sequence Registration (BraTS-Reg) Challenge"
doi: 10.5281/zenodo.6362419
[8] Kickingereder, P., Isensee, F. et al. "Automated quantitative tumor response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study." The Lancet Oncology 20, 728-740 (2019)
https://doi.org/10.1016/S1470-2045(19)30098-1
[9] Suter, Y., Knecht, U., Valenzuela, W., Notter, M., Hewer, E., Schucht, P., Wiest, R. and Reyes, M., 2022. "The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation." Scientific data, 9(1), p.768.
[10] Isensee, F., Jaeger, P.F., Kohl, S.A.A. et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature Methods 18, 203-211 (2021).
https://doi.org/10.1038/s41592-020-01008-z
[11] HD-GLIO-AUTO: https://github.com/NeuroAI-HD/HD-GLIO-AUTO
[12] DeepBraTumIA: https://www.nitrc.org/projects/deepbratumia/
[13] Isensee, F., Schell, M., Tursunova, I. et al. "Automated brain extraction of multi-sequence MRI using artificial neural networks." Human Brain Mapping 40, 4952-4964 (2019)
https://doi.org/10.1002/hbm.24750
[14] Maier-Hein, L., Reinke, A. et al. "Metrics reloaded: Pitfalls and recommendations for image analysis validation" ArXiv:2206.01653 [Cs] (2022)
https://doi.org/10.48550/arxiv.2206.01653
[15] Maier-Hein, L., Eisenmann, M., Reinke, A. et al. "Why rankings of biomedical image analysis competitions should be interpreted with care" Nature Communications 9, 5217 (2018).
https://doi.org/10.1038/s41467-018-07619-7

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