MICCAI Brain Tumor Segmentation (BraTS) 2020 Benchmark: "Prediction of Survival and Pseudoprogression": Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

MICCAI Brain Tumor Segmentation (BraTS) 2020 Benchmark: "Prediction of Survival and Pseudoprogression"

Challenge acronym

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

BraTS 2020

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.

BraTS 2020 utilizes multi-institutional MRI scans and focuses on the segmentation of intrinsically heterogeneous (in appearance, shape, and histology) brain tumors, namely gliomas. Compared to BraTS'17-'19, this year BraTS includes both pre-operative and post-operative scans (i.e., including surgically imposed cavities) and attempts to quantify the uncertainty of the predicted segmentations. Furthermore, to pinpoint the clinical relevance of the segmentation task, BraTS'20 also focuses on 1) the prediction of patient overall survival from pre-operative scans (Task 2) and 2) the distinction between true tumor recurrence and treatment related effects on the post-operative scans (Task 3), via integrative analyses of quantitative imaging phenomic features and machine learning algorithms. Ground truth annotations are created and approved by expert neuroradiologists for every subject included in the training, validation, and testing datasets to quantitatively evaluate the predicted tumor segmentations (Task 1). Furthermore, the quantitative evaluation of the clinically-relevant tasks (i.e., overall survival (Task 2) and distinction between tumor recurrence and treatment related effects (Task 3)), is performed according to real clinical data. Participants are free to choose whether they want to focus only on one or multiple tasks.

Challenge keywords

List the primary keywords that characterize the challenge.

brain tumor, segmentation, glioblastoma, glioma, uncertainty, survival prediction, pseudoprogression, recurrence

Year

The challenge will take place in ...

2020

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

Brain Lesion (BrainLes) workshop 2020.

Duration

How long does the challenge take?

Half day.

Expected number of participants

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

We conservatively estimate participation from (at least) 50 teams, considering:

i) the continuously increasing number of participating teams during the past 7 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),

ii) last year we had 637 requests to download the training data and 72 participating teams who submitted results on the final testing phase,

iii) since Oct 2019, that BraTS took place in MICCAI 2019 we have 611 additional requests for downloading the BraTS 2019 data, and we expect these users to be interested in participating in BraTS 2020,

iv) we have already received 52 explicit requests to download the BraTS 2020 training data when available, and emails of intention to participate in BraTS 2020 from 16 international groups.

In addition, we will advertise the event in related mailing lists (e.g., CVML) and we intend to send an email to all the above and notify them about this year's challenge.

Publication and future plans

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

The configuration of combining the BraTS challenge with the BrainLes workshop provides the BraTS participants with the option to extend their papers to 12 pages, and hence publish their methods in the workshop's LNCS post-conference proceedings.

Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of BraTS 2020, making comparative assessment with the summary results of the previous BraTS challenges, in particular focusing on the evaluation of segmentation uncertainty and evaluating the effect of varying segmentation labels in research beyond segmentation, e.g., radiomic analyses. The main focus of this publication will be on whether methods are harshly penalized in areas that even clinical experts are uncertain.

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.

BraTS is an off-site challenge and algorithms are run using the participants' computing infrastructure.

Hardware requirements: 1 (or 2) projectors, 2 microphones, loudspeakers

TASK: Segmentation of gliomas in preoperative MRI scans

SUMMARY

Keywords

List the primary keywords that characterize the task.

segmentation

ORGANIZATION

Organizers

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

Organizing team:

Spyridon Bakas, Ph.D., – [Lead Organizer - Contact Person]
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Bjoern Menze, Ph.D.,
Technical University of Munich (TUM), Germany

Christos Davatzikos, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Jayashree Kalpathy-Cramer, Ph.D.,
Athinoula A. Martinos for Biomedical Imaging, Massachusetts General Hospital (MGH), Harvard Medical School, USA

Keyvan Farahani, Ph.D.,
Cancer Imaging Program, National Cancer Institute (NCI), National Institutes of Health (NIH), USA

Clinical Evaluators and Annotation Approvers:

Michel Bilello, MD, Ph.D.,
University of Pennsylvania, Philadelphia, PA, USA

Suyash Mohan, MD, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Data Contributors:

John B. Freymann & Justin S. Kirby - on behalf of The Cancer Imaging Archive (TCIA), Cancer Imaging Program,

NCI, National Institutes of Health (NIH), USA

Christos Davatzikos, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Manmeet Ahluwalia, M.D. & Volodymyr Statsevych, M.D.,
Cleveland Clinic, , Cleveland, OH, USA

Raymond Huang, M.D., Ph.D.,
Brigham and Women's Hospital, Boston, MA, USA

Hassan Fathallah-Shaykh, M.D., Ph.D.,
University of Alabama at Birmingham, AL, USA

Roland Wiest, M.D.,
University of Bern, Switzerland

Andras Jakab, M.D., Ph.D.,
University of Debrecen, Hungary

Rivka R. Colen, M.D.
University of Pittsburgh Medical Center

Aikaterini Kotrotsou, Ph.D.,
MD Anderson Cancer Center, TX, USA

Daniel Marcus, Ph.D., & Mikhail Milchenko, Ph.D., & Arash Nazeri, M.D.,
Washington University School of Medicine in St.Louis, MO, USA

Marc-Andre Weber, M.D.,
Heidelberg University, Germany

Abhishek Mahajan, M.D. & Ujjwal Baid, Ph.D.,
Tata Memorial Center, Mumbai, India

b) Provide information on the primary contact person.

Spyridon Bakas, Ph.D., – [Lead Organizer - Contact Person]
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA
brats2020@cbica.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.

Examples:

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

Repeated event one time.

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.

In consistency with the last 3 years, we will be using the University of Pennsylvania's Image Processing Portal (ipp.cbica.upenn.edu) for running the challenge.

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

https://www.med.upenn.edu/cbica/brats2020.html

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 data augmentation, but if they do so, they MUST also discuss the potential difference in their results after using only the BraTS 2020 data, since our intention is to solve the brain tumor 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 not eligible for awards and not listed in leaderboard.

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

We expect that, similarly to BraTS 2018-2019, Intel AI will sponsor a $5K award for the top 3 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.

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 combining the BraTS challenge with the BrainLes workshop provides the BraTS participants with the option to extend their papers to 12 pages, and hence publish their methods in the workshop's LNCS post-conference proceedings.

Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of BraTS 2020, making comparative assessment with the summary results of the previous BraTS challenges.

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 algorithms to the evaluation platform for the scoring to occur. Furthermore, for the final testing phase the participants will be requested to submit their algorithm in the form of a docker container to the "BraTS Algorithmic Repository" as described in: https://www.med.upenn.edu/sbia/brats2018/algorithms.html

This will enable confirmation of reproducibility, running of these algorithms to the previous BraTS instances and comparison with results obtained by algorithms of previous years, thereby maximization of benefit towards solving the problem of brain tumor segmentation.

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 a validation set in June, allowing participants to tune their methods in unseen data. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform (ipp.cbica.upenn.edu) will be allowed.

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 challenge's approval until submission deadline of short papers reporting method and preliminary results (see below).
Expected release of training data: 01 May 2020
Expected release of validation data: 26 June 2020
Submission of short papers, reporting method & preliminary results: 25 July 2020.

Expected release of testing data & evaluation within 48hrs. (Only for participants with submitted papers): 17-28 August 2020
Contacting top performing methods for preparing slides for oral presentation: 4 Sep 2020
Announcement of final top 3 ranked teams: Challenge at MICCAI

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 intend to make the data of the MICCAI BraTS 2020 challenge available via The Cancer Imaging Archive (TCIA) of the National Institutes of Health (NIH), following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions).

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: If any of the non-TCIA contributors object to this license, the specific subset of the BraTS 2020 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 ranking code will be available after the end of the challenge.

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

This year the participants are required to submit their dockerized algorithm together with their final submitted results, during the testing phase. Specific instructions for creating docker containers of uniform API for the BraTS challenge will be provided similar to the “BraTS Algorithmic Repository” described in: https://www.med.upenn.edu/sbia/brats2018/algorithms.html

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 AI.

Spyridon Bakas and the clinical evaluators will have access to the 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, Assistance, Treatment planning, Intervention planning, Longitudinal study, Surgery, Training, Diagnosis, CAD.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 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).

N/A

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

Age, Gender, Resection Status, Progression Status.

Target entity(ies)

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

Brain mpMRI scans.

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

Tumor in the brain.

Assessment aim(s)

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

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

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

Specificity, Sensitivity.

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 the related Nature Scientific Data manuscript of ours [1]. Since then, multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 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 Nature Scientific Data manuscript of ours [1]. Since then multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 pre-operative multimodal MRI 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) MD Anderson Cancer Center (TX, USA),
18) Washington University in St. Louis (MO, USA),
19) Tata Memorial Center (India),
20) University of Pittsburg Medical Center (PA, USA),
21) Cleveland Clinic (OH, USA),
22) Brigham and Women's Hospital (MA, USA).

Note that data from institutions 6-16 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).

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

Training and test case characteristics

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

Examples:

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

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

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

BraTS 2020 training, validation, and testing data will be an extended dataset configuration since BraTS'19.
The exact numbers at the very minimum will be:
Training data: 420 patients
Validation data: 150 patients

Testing data: 230 patients

For further data augmentation we already 1) working with clinical experts from our institutions to manually annotate scans of existing publicly available datasets (e.g., IvyGAP, ECOG-ACRIN clinical trial available data available through the TCIA), and 2) provide more multi-parametric MRI scans of gliomas from our affiliated institutions.

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

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

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 From Multiple Human Raters (>2).

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

Annotators were given specific instructions on what the segmentations of the specific tumor sub-regions should describe. As long as an experienced neuroradiologist is approving the final ground truth segmentation labels, the annotators were given the flexibility to use either a manual annotation approach, or a hybrid approach where an automated method is used to produce some bulk annotations followed by manual refinements.
Summary of specific instructions.
i) the farthest tumor extent including the edema (what is called the whole tumor), delineates the hyperintense regions with homogeneous signal on T2 & T2-FLAIR.
ii) the tumor core (including the enhancing, non-enhancing, and necrotic tumor) delineates regions of lower T2 signal.
iii) the enhancing tumor delineates the hyperintense signal of the T1-Gd, after excluding the vessels.
iv) the necrotic core outlines regions appearing dark in both T1 and T1-Gd images (denoting necrosis/cysts), and dark regions in T1-Gd and bright in T1.

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.

At least 2 experienced neuroradiologists with >12 years of experience.

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.

MRI scans have been co-registered to the same anatomical template (i.e., SRI atlas [2]), interpolated to the same resolution (1 mm^3) and skull-stripped following manual revision.

[2] https://www.ncbi.nlm.nih.gov/pubmed/20017133

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 the introduction of the Task 4, i.e., to quantify the uncertainty in the tumor segmentations.

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

N/A

ASSESSMENT METHODS

Metric(s)

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

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

Dice Similarity Coefficient (DSC),
95% Hausdorff distance (HD),
Sensitivity,
Specificity,
The regions evaluated using these metrics describe the whole tumor, the tumor core, and the enhancing tumor. 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 extend of resection.
ii) the tumor core describes what is typically resected during a surgical procedure.
iii) the whole tumor as it defines the whole extent of the tumor, including the peritumoral edematous tissue and

highly infiltrated area.

In terms of evaluation metrics we use:
i) the Dice Similarity Coefficient, which is commonly used in the assessment of segmentation performance,
ii) the 95% Hausdorff distance as opposed to standard HD, in order to avoid outliers havings too much weight,
iii) Sensitivity and Specificity to determine whether an algorithm has the tendency to over- or undersegment.

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.

Similarly to BraTS 2017 - BraTS 2019, each participant will be ranked for each of the X test cases. Each case includes 3 regions of evaluation, and the metrics used to produce the rankings will be the Dice Similarity Coefficient and the 95% Hausdorff distance. Thus, for X number of cases included in the BraTS 2020, each participant ends up having X*3*2 rankings. The final ranking score is the average of all these rankings normalized by the number of teams.

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

Missing results on test cases, will result in the ranks for the corresponding metrics to be set to the maximum.

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

Following discussions with our biostatistician, 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.

Uncertainties in rankings will be assessed using permutational analyses. 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, and uncertainty in these final ranks will be measured by permuting the relative ranks for each segmentation measure and tissue class.

To assess statistically significant differences, we will use paired and unpaired rank-based and t-test statistics for errors compared with permutation-generated one-sided null distributions..

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.

TASK: Survival Prediction

SUMMARY

Keywords

List the primary keywords that characterize the task.

survival prediction

ORGANIZATION

Organizers

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

Organizing team:

Spyridon Bakas, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Bjoern Menze, Ph.D.,
Technical University of Munich (TUM), Germany

Russell Taki Shinohara, Ph.D.,
Penn Statistics in Imaging and Visualization Endeavor (PennSIVE), University of Pennsylvania, Philadelphia, PA, USA

Christos Davatzikos, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Clinical Evaluators and Annotation Approvers:

Michel Bilello, MD, Ph.D.,
University of Pennsylvania, Philadelphia, PA, USA

Suyash Mohan, MD, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Data Contributors:

John B. Freymann & Justin S. Kirby - on behalf of The Cancer Imaging Archive (TCIA), Cancer Imaging Program, NCI, National Institutes of Health (NIH), USA

Christos Davatzikos, Ph.D.,

Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Hassan Fathallah-Shaykh, M.D., Ph.D.,
University of Alabama at Birmingham, AL, USA

Roland Wiest, M.D.,
University of Bern, Switzerland

Rivka R. Colen, M.D.
University of Pittsburgh Medical Center

Aikaterini Kotrotsou, Ph.D.,
MD Anderson Cancer Center, TX, USA

Daniel Marcus, Ph.D., & Mikhail Milchenko, Ph.D., & Arash Nazeri, M.D.,
Washington University School of Medicine in St.Louis, MO, USA

Marc-Andre Weber, M.D.,
Heidelberg University, Germany

Abhishek Mahajan, M.D., & Ujjwal Baid, Ph.D.,
Tata Memorial Center, Mumbai, India

b) Provide information on the primary contact person.

Spyridon Bakas, Ph.D., – [Lead Organizer - Contact Person]
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA
brats2020@cbica.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.

Examples:

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

Repeated event one time.

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.

In consistency with the last 3 years, we will be using the University of Pennsylvania's Image Processing Portal (ipp.cbica.upenn.edu) for running the challenge.

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

https://www.med.upenn.edu/cbica/brats2020.html

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 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.

May participate but not eligible for awards and not listed in leaderboard.

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

No prize included.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

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

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

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

The configuration of combining the BraTS challenge with the BrainLes workshop provides the BraTS participants with the option to extend their papers to 12 pages, and hence publish their methods in the workshop's LNCS post-conference proceedings.

Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of BraTS 2020 making comparative assessment with the summary results of the previous BraTS challenges, and evaluating the effect of varying segmentation labels in research beyond segmentation, i.e., survival prediction.

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 algorithms to the evaluation platform for the scoring to

occur. Furthermore, for the final testing phase the participants will be requested to submit their algorithm in the form of a docker container to the "BraTS Algorithmic Repository" as described in: https://www.med.upenn.edu/sbia/brats2018/algorithms.html

This will enable confirmation of reproducibility, running of these algorithms to the previous BraTS instances and comparison with results obtained by algorithms of previous years.

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 a validation set in June, allowing participants to tune their methods in unseen data. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform (ipp.cbica.upenn.edu) will be allowed.

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 challenge's approval until submission deadline of short papers reporting method and preliminary results (see below).
Expected release of training data: 01 May 2020
Expected release of validation data: 26 June 2020
Submission of short papers, reporting method & preliminary results: 25 July 2020.
Expected release of testing data & evaluation within 48hrs. (Only for participants with submitted papers): 17-28 August 2020
Contacting top performing methods for preparing slides for oral presentation: 4 Sep 2020
Announcement of final top 3 ranked teams: Challenge at MICCAI.

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 intend to make the data of the MICCAI BraTS 2020 challenge available via The Cancer Imaging Archive (TCIA) of the National Institutes of Health (NIH), following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions).

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: If any of the non-TCIA contributors object to this license, the specific subset of the BraTS 2020 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 ranking code will be available after the end of the challenge.

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

The participants of this task will be asked to submit their docker together with their final submitted results, during the testing phase. Specific instructions for creating docker containers of uniform API for the BraTS challenge will be provided similar to the “BraTS Algorithmic Repository” described in:
https://www.med.upenn.edu/sbia/brats2018/algorithms.html

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.

N/A

Spyridon Bakas and the clinical evaluators will have access to the 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, Assistance, Longitudinal study, Treatment planning, Intervention planning, Surgery, Training, Screening, Diagnosis, Decision support, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

Cohorts

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

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with

multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Imaging modality(ies)

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

MRI + clinical info.

Context information

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

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

N/A

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

Age, Gender, Resection Status, Progression Status.

Target entity(ies)

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

Brain mpMRI scans.

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

Tumor in the brain.

Assessment aim(s)

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

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

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

Accuracy.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

The exact scanners and their technical specifications used for acquiring the TCIA cohort has been listed in the data reference published in the related Nature Scientific Data manuscript of ours [1]. Since then, multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 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 Nature Scientific Data manuscript of ours [1]. Since then multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 pre-operative multimodal MRI 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) MD Anderson Cancer Center (TX, USA),
18) Washington University in St. Louis (MO, USA),

19) Tata Memorial Center (India),
20) University of Pittsburg Medical Center (PA, USA),
21) Cleveland Clinic (OH, USA),
22) Brigham and Women's Hospital (MA, USA).

Note that data from institutions 6-16 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).

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

Training and test case characteristics

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

Examples:

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

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

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

BraTS 2020 training, validation, and testing data will be an extended dataset configuration since BraTS'19.
The exact numbers at the very minimum will be:
Training data: 420 patients
Validation data: 150 patients
Testing data: 230 patients
A subset of these will be included in the survival prediction task, depending on which of these have corresponding survival information.

For further data augmentation we already 1) working with clinical experts from our institutions to manually annotate scans of existing publicly available datasets (e.g., IvyGAP, ECOG-ACRIN clinical trial available data available through the TCIA), and 2) provide more multi-parametric MRI scans of gliomas from our affiliated institutions.

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

Based on availability. The data was split in these numbers between training, validation, and testing after

considering the number of cases used as test cases in previous instances of BraTS and the fact that the organizers did not want to reveal ground truth labels of previous test cases.

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 was chosen according to real-world distribution.

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.

Overall survival provided by the patient records.

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.

N/A

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.

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.

N/A

Data pre-processing method(s)

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

MRI scans have been co-registered to the same anatomical template (i.e., SRI atlas [2]), interpolated to the same resolution (1 mm^3) and skull-stripped following manual revision.

[2] https://www.ncbi.nlm.nih.gov/pubmed/20017133

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

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, MSE

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

We use:
i) the Accuracy, to evaluate the number of correctly classified survivors over all patients,
ii) the Mean Squared Error and the Median Standard Deviation, to assess outliers within classes and any pairwise error between predicted and actual survival.

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.

Similarly to BraTS 2017 - BraTS 2019, each participant will be ranked based on the obtained accuracy.

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

Missing results on test cases, will result in the ranks for the corresponding metrics to be set to the maximum.

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

Following discussions with our biostatistician, 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.

We will analyze uncertainty in rankings by permuting estimated survival times across methods and comparing error distributions.

To assess statistically significant differences, we will use paired and unpaired rank-based and t-test statistics for errors compared with permutation-generated one-sided null distributions.

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.

TASK: Distiction of Tumor Recurrence from Treatment Related Effects

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.

One of the most challenging clinical problems in management of gliomas is distinguishing pseudo-progression (PsP), a benign side-effect of the aggressive chemoradiation therapy, from tumor recurrence. The current standard-of-care for evaluating treatment response in brain tumors involves follow-up imaging examinations using T1w, T2w and T2-FLAIR MRI scans. Unfortunately, guidelines set by the clinical Response Assessment in Neuro-Oncology (RANO) criteria [3] are based solely on bi-directional diametric measurements of enhancement observed on T1w, T2w/FLAIR scans and sub-optimal often leading to unnecessary surgical interventions for disease confirmation in patients with a benign condition. In this task of the BraTS 2020 challenge, we will release a well-curated cohort of multi-institutional retrospective studies with the expected goal that the participating teams will develop radiomics and machine learning solutions to be able to reliably distinguish benign PsP from tumor recurrence using routinely acquired standard-of-care MRI scans.

[3] P.Y.Wen, D.R.Macdonald, D.A.Reardon, et al., "Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group." J Clin Oncol. 28:1963-1972, 2010.
Publication link: https://ascopubs.org/doi/full/10.1200/JCO.2009.26.3541

Keywords

List the primary keywords that characterize the task.

pseudoprogression, classification

ORGANIZATION

Organizers

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

Organizing team:

Pallavi Tiwari, Ph.D.,
Case Western Reserve University, Cleveland, OH, USA

Spyridon Bakas, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Bjoern Menze, Ph.D.,
Technical University of Munich (TUM), Germany

Christos Davatzikos, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Clinical Evaluators and Annotation Approvers:

Raymond Huang, M.D., Ph.D.,
Brigham and Women's Hospital, Boston, MA, USA

Manmeet Ahluwalia, M.D. & Volodymyr Statsevych, M.D.,
Cleveland Clinic, Cleveland, OH, USA

Michel Bilello, MD, Ph.D.,
University of Pennsylvania, Philadelphia, PA, USA

Suyash Mohan, MD, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Data Contributors:

John B. Freymann & Justin S. Kirby - on behalf of The Cancer Imaging Archive (TCIA), Cancer Imaging Program, NCI, National Institutes of Health (NIH), USA

Christos Davatzikos, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Manmeet Ahluwalia, M.D. & Volodymyr Statsevych, M.D.,
Cleveland Clinic, Cleveland, OH, USA

Raymond Huang, M.D., Ph.D.,
Brigham and Women's Hospital, Boston, MA, USA

b) Provide information on the primary contact person.

Spyridon Bakas, Ph.D., – [Lead Organizer - Contact Person]
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA
brats2020@cbica.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.

Examples:

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

One time event.

Challenge venue and platform

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

MICCAI.

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

We will be using the University of Pennsylvania's Image Processing Portal (ipp.cbica.upenn.edu) for running the challenge.

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

https://www.med.upenn.edu/cbica/brats2020.html

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 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.

May participate but not eligible for awards and not listed in leaderboard.

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

No prize included.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

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

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

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

The configuration of combining the BraTS challenge with the BrainLes workshop provides the BraTS participants

with the option to extend their papers to 12 pages, and hence publish their methods in the workshop's LNCS post-conference proceedings.

Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of BraTS 2020, making comparative assessment with the summary results of the previous BraTS challenges, and evaluating the effect of varying segmentation labels in research beyond segmentation, i.e., distinguishing pseudoprogression from true recurrence.

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 algorithms to the evaluation platform for the scoring to occur. Furthermore, for the final testing phase the participants will be requested to submit their algorithm in the form of a docker container to the "BraTS Algorithmic Repository" as described in: https://www.med.upenn.edu/sbia/brats2018/algorithms.html

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 a validation set in June, allowing participants to tune their methods in unseen data. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform (ipp.cbica.upenn.edu) will be allowed.

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 challenge's approval until submission deadline of short papers reporting method and preliminary results (see below).
Expected release of training data: 01 May 2020
Expected release of validation data: 26 June 2020
Submission of short papers, reporting method & preliminary results: 25 July 2020.
Expected release of testing data & evaluation within 48hrs. (Only for participants with submitted papers): 17-28 August 2020
Contacting top performing methods for preparing slides for oral presentation: 4 Sep 2020
Announcement of final top 3 ranked teams: Challenge at MICCAI

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 intend to make the data of the MICCAI BraTS 2020 challenge available via The Cancer Imaging Archive (TCIA) of the National Institutes of Health (NIH), following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions).

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: If any of the non-TCIA contributors object to this license, the specific subset of the BraTS 2020 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 ranking code will be available after the end of the challenge.

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

The participants of this task will be asked to submit their docker together with their final submitted results, during the testing phase. Specific instructions for creating docker containers of uniform API for the BraTS challenge will be provided similar to the “BraTS Algorithmic Repository” described in: https://www.med.upenn.edu/sbia/brats2018/algorithms.html

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.

N/A

Spyridon Bakas, Pallavi Tiwari, and the clinical evaluators will have access to the 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, Assistance, Longitudinal study, Treatment planning, Intervention planning, Surgery, Training, Screening, Diagnosis, Decision support, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

Cohorts

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

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

Imaging modality(ies)

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

MRI + clinical info.

Context information

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

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

N/A

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

Age, Gender, Resection Status, Progression Status.

Target entity(ies)

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

Brain mpMRI scans.

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

Tumor in the brain.

Assessment aim(s)

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

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

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

Accuracy.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

The exact scanners and their technical specifications used for acquiring the TCIA cohort has been listed in the data reference published in the related Nature Scientific Data manuscript of ours [1]. Since then, multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 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 Nature Scientific Data manuscript of ours [1]. Since then multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 pre-operative multimodal MRI 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) MD Anderson Cancer Center (TX, USA),
18) Washington University in St. Louis (MO, USA),
19) Tata Memorial Center (India),
20) University of Pittsburg Medical Center (PA, USA),
21) Cleveland Clinic (OH, USA),
22) Brigham and Women's Hospital (MA, USA).

Note that data from institutions 6-16 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).

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

Training and test case characteristics

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

Examples:

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

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

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

The exact numbers at the very minimum will be:
Training data: 80 patients
Validation data: 20 patients
Testing data: 40 patients

For further data augmentation we already 1) working with clinical experts from our institutions to manually annotate scans of existing publicly available datasets (e.g., IvyGAP, ECOG-ACRIN clinical trial available data available through the TCIA), and 2) provide more multi-parametric MRI scans of gliomas from our affiliated institutions.

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.

Class distribution was chosen according to real-world distribution.

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.

Progression status following tissue evaluation from expert neuropathologists.

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.

N/A

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.

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.

N/A

Data pre-processing method(s)

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

MRI scans have been co-registered to the same anatomical template (i.e., SRI atlas [2]), interpolated to the same resolution (1 mm^3) and skull-stripped following manual revision.

[2] https://www.ncbi.nlm.nih.gov/pubmed/20017133

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.

None, other the tissue sampling error.

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,
Sensitivity,
Specificity.

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

We use:
i) the Accuracy, to evaluate the number of correct classified progression status,
ii) the sensitivity, as it is clinically important to be able to reliably identify tumor recurrence cases.

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.

Sensitivity, specificity, and overall accuracy will be used as evaluation metrics. While it is important to have improved accuracies, emphasis will also be given to high sensitivity because clinically it is more important to be able to reliably identify tumor recurrence cases.
Each participant will be ranked based on the obtained accuracy.

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

Missing results on test cases, will result in the ranks for the corresponding metrics to be set to the maximum.

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

Following discussions with our biostatistician, 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.

We will analyze uncertainty in rankings by permuting estimated progression status across methods and comparing error distributions.

To assess statistically significant differences, we will use paired and unpaired rank-based and t-test statistics for

errors compared with permutation-generated one-sided null distributions.

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.

TASK: Uncertainty Quantification

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.

Although many machine learning methods have been developed for brain tumor segmentation, errors in segmentation still persist. Producing uncertainties in the resulting segmentation results will build trust by clinicians in the results of deep learning models, and will therefore encourage integration of the resulting models into real clinical practice. However, to this end, the uncertainties produced need to properly reflect the accuracy of the system. This task of the BraTS 2020 challenge seeks to reward uncertainties that are confident when correct, and are uncertain when incorrect. This will permit more appropriate clinical evaluation of the results and potential integration into downstream methods. There is currently no clear benchmark available which permits accurate comparisons between uncertainty results generated by different methods in this context. In MICCAI 2019, we ran a preliminary experimental task during BraTS 2019 to assess several benchmark approaches. With growing excitement for the research area, the BraTS 2019 uncertainty task attracted 15 teams. The methods were ranked according to scores linked to "DICE" provided by the main challenge. We refined the metrics based on the validation results. The BraTS 2019 uncertainty task was a success and the benchmarks reflected the objectives. This year we rerun the same uncertainty task based on the same metrics, with additional ones to be added that reflect voxel based segmentation results in addition to "DICE" scores, in order to further assess the various uncertainty measures developed in the field. The objective of the BraTS 2020 uncertainty task is to provide a stable benchmark for the quantification of uncertainty for the context of brain tumor segmentation.

Keywords

List the primary keywords that characterize the task.

uncertainty quantification

ORGANIZATION

Organizers

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

Organizing team:

Tal Arbel, Ph.D., & Raghav Mehta
McGill University, Canada

Spyridon Bakas, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Yarin Gal, Ph.D., & Angelos Filos
University of Oxford, UK

Bjoern Menze, Ph.D.,
Technical University of Munich (TUM), Germany

Clinical Evaluators and Annotation Approvers:

Michel Bilello, MD, Ph.D.,
University of Pennsylvania, Philadelphia, PA, USA

Suyash Mohan, MD, Ph.D.
University of Pennsylvania, Philadelphia, PA, USA

Data Contributors:

John B. Freymann & Justin S. Kirby - on behalf of The Cancer Imaging Archive (TCIA), Cancer Imaging Program, NCI, National Institutes of Health (NIH), USA

Christos Davatzikos, Ph.D.,
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

Manmeet Ahluwalia, M.D. & Volodymyr Statsevych, M.D.,
Cleveland Clinic, , Cleveland, OH, USA

Raymond Huang, M.D., Ph.D.,
Brigham and Women's Hospital, Boston, MA, USA

Hassan Fathallah-Shaykh, M.D., Ph.D.,
University of Alabama at Birmingham, AL, USA

Roland Wiest, M.D.,
University of Bern, Switzerland

Andras Jakab, M.D., Ph.D.,
University of Debrecen, Hungary

Rivka R. Colen, M.D. & Aikaterini Kotrotsou, Ph.D.,
MD Anderson Cancer Center, TX, USA

Rivka R. Colen, M.D.
University of Pittsburgh Medical Center

Daniel Marcus, Ph.D., & Mikhail Milchenko, Ph.D., & Arash Nazeri, M.D.,

Washington University School of Medicine in St.Louis, MO, USA

Marc-Andre Weber, M.D.,
Heidelberg University, Germany

Abhishek Mahajan, M.D., & Ujjwal Baid, Ph.D.,
Tata Memorial Center, Mumbai, India

b) Provide information on the primary contact person.

Spyridon Bakas, Ph.D., – [Lead Organizer - Contact Person]
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA
brats2020@cbica.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.

Examples:

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

Repeated event one time.

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.

In consistency with the last year, we will be using the University of Pennsylvania's Image Processing Portal (ipp.cbica.upenn.edu) for running the challenge.

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

https://www.med.upenn.edu/cbica/brats2020.html

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 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.

May participate but not eligible for awards and not listed in leaderboard.

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

No prize included.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

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

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

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

The configuration of combining the BraTS challenge with the BrainLes workshop provides the BraTS participants with the option to extend their papers to 12 pages, and hence publish their methods in the workshop's LNCS post-conference proceedings.

Furthermore, we intend to coordinate a journal manuscript focusing on publishing and summarizing the results of BraTS 2020, making comparative assessment with the summary results of the BraTS 2019 uncertainty task.

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 algorithms to the evaluation platform for the scoring to occur. Furthermore, for the final testing phase the participants will be requested to submit their algorithm in the form of a docker container to the "BraTS Algorithmic Repository" as described in: https://www.med.upenn.edu/sbia/brats2018/algorithms.html

This will enable confirmation of reproducibility, running of these algorithms to the previous BraTS instances and comparison with results obtained by algorithms of previous years.

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 a validation set in June, allowing participants to tune their methods in unseen data. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform (ipp.cbica.upenn.edu) will be allowed.

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 challenge's approval until submission deadline of short papers reporting method and preliminary results (see below).
Expected release of training data: 01 May 2020
Expected release of validation data: 26 June 2020
Submission of short papers, reporting method & preliminary results: 25 July 2020.
Expected release of testing data & evaluation within 48hrs. (Only for participants with submitted papers): 17-28 August 2020
Contacting top performing methods for preparing slides for oral presentation: 4 Sep 2020
Announcement of final top 3 ranked teams: Challenge at MICCAI

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 intend to make the data of the MICCAI BraTS 2020 challenge available via The Cancer Imaging Archive (TCIA) of the National Institutes of Health (NIH), following their standard licensing (https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions).

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: If any of the non-TCIA contributors object to this license, the specific subset of the BraTS 2020 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 ranking code will be available after the end of the challenge.

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

The participants of this task will be asked to submit their docker together with their final submitted results, during the testing phase. Specific instructions for creating docker containers of uniform API for the BraTS challenge will be provided similar to the "BraTS Algorithmic Repository" described in: https://www.med.upenn.edu/sbia/brats2018/algorithms.html

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.

N/A

Spyridon Bakas and the clinical evaluators will have access to the 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, Assistance, Treatment planning, Training, Diagnosis, Decision support.

Task category(ies)

State the task category(ies).

Examples:

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

Uncertainty Quantification.

Cohorts

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

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) MRI.

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

Retrospective multi-institutional cohort of patients, diagnosed with de novo glioma tumors, clinically scanned with multi-parametric MRI acquisition protocol including i) native and ii) contrast-enhanced T1-weighted, iii) T2-weighted and iv) T2 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).

N/A

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

Age, Gender, Resection Status, Progression Status.

Target entity(ies)

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

Brain mpMRI scans.

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

Tumor in the brain.

Assessment aim(s)

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

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

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

Accuracy.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

The exact scanners and their technical specifications used for acquiring the TCIA cohort has been listed in the data reference published in the related Nature Scientific Data manuscript of ours [1]. Since then, multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 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 Nature Scientific Data manuscript of ours [1]. Since then multiple institutions have contributed data to the create the current BraTS dataset. We are currently in coordination with TCIA to make the BraTS data permanently available through them. All the acquisition details will be included together with the data availability in TCIA.

[1] 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 pre-operative multimodal MRI 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) MD Anderson Cancer Center (TX, USA),
18) Washington University in St. Louis (MO, USA),
19) Tata Memorial Center (India),
20) University of Pittsburg Medical Center (PA, USA),
21) Cleveland Clinic (OH, USA),
22) Brigham and Women's Hospital (MA, USA).

Note that data from institutions 6-16 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).

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

Training and test case characteristics

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

Examples:

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

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

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

BraTS 2020 training, validation, and testing data will be an extended dataset configuration since BraTS'19.
The exact numbers at the very minimum will be:
Training data: 420 patients
Validation data: 150 patients
Testing data: 230 patients

For further data augmentation we already 1) working with clinical experts from our institutions to manually annotate scans of existing publicly available datasets (e.g., IvyGAP, ECOG-ACRIN clinical trial available data available through the TCIA), and 2) provide more multi-parametric MRI scans of gliomas from our affiliated institutions.

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

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

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 From Multiple Human Raters (>2).

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

Annotators were given specific instructions on what the segmentations of the specific tumor sub-regions should describe. As long as an experienced neuroradiologist is approving the final ground truth segmentation labels, the annotators were given the flexibility to use either a manual annotation approach, or a hybrid approach where an automated method is used to produce some bulk annotations followed by manual refinements.
Summary of specific instructions.
i) the farthest tumor extent including the edema (what is called the whole tumor), delineates the hyperintense regions with homogeneous signal on T2 & T2-FLAIR.
ii) the tumor core (including the enhancing, non-enhancing, and necrotic tumor) delineates regions of lower T2 signal.
iii) the enhancing tumor delineates the hyperintense signal of the T1-Gd, after excluding the vessels.
iv) the necrotic core outlines regions appearing dark in both T1 and T1-Gd images (denoting necrosis/cysts), and dark regions in T1-Gd and bright in T1.s

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.

At least 2 experienced neuroradiologists with >12 years of experience.

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.

MRI scans have been co-registered to the same anatomical template (i.e., SRI atlas [2]), interpolated to the same resolution (1 mm^3) and skull-stripped following manual revision.

[2] https://www.ncbi.nlm.nih.gov/pubmed/20017133

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 the introduction of the Task 4, i.e., to quantify the uncertainty in the tumor segmentations.

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)

For the task of estimating uncertainty, uncertain voxels will be filtered out at several predetermined N number of uncertainty threshold points "Thr", and the model performance will be assessed based on the "Dice" of the remaining voxels at each of these Thr. For example, Thr:75 implies that all voxels with uncertainty values >75 will be marked as uncertain and the associated predictions will be filtered out and not considered for the subsequent Dice calculations. Dice values will only be calculated for the remaining predictions at the unfiltered voxels. This evaluation will reward approaches where the confidence in the correct assertions is high (True Positives - TPs / True Negatives - TNs) and low for incorrect assertions (False Positives - FPs, and False Negatives - FNs). For these approaches, it is expected that as more uncertain voxels are filtered out, the Dice score will increase on the remaining predictions. A second evaluation will keep track of the ratio of TPs/TNs that are filtered relative to the initial/baseline number of TPs/TNs (TP/TN at threshold 100) at different Thr. This evaluation will essentially penalize approaches that filter out a large percentage of TP/TN voxels, in order to attain the reported Dice value, and thereby rewarding approaches with a lower percentage of uncertain TPs/TNs.

It should be noted that, as there is currently no specific metric available for this task in the literature, we are still exploring different metrics instead of just using "Dice". We may change the final metrics before the training dataset becomes publicly available.

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

There is currently, no specific metric available in the literature which will allows us to compare uncertainty generated by different methods. We designed the above mentioned metric, keeping in mind following criterion. For a Computer-Aided Diagnosis (CAD) system of pathology segmentation where the size of the pathology is tiny compared to its surrounding healthy tissues, as in the case of brain tumor segmentation, we want the system to be (a) confident when correct and (b) uncertain when incorrect. This will allow the output of an automatic segmentation system where the system is wrong to be flagged and corrected by an experienced clinician without putting overburden on the clinician by flagging output where the system is correct.

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 ranking system of the BraTS 2019 sub-challenge and compute AUC curves for three different curves: 1) Dice vs Uncertainty threshold 2) FTP vs Uncertainty threshold 3) FTN vs Uncertainty threshold. We will combine these AUCs to generate a score in the following manner: score = AUC_1 + (1-AUC-2) + (1-AUC_3). We will be experimenting with using AUC of Precision-Recall vs Uncertainty thresholds and either integrate with the score above or have a separate scoring method. We will generate above mentioned score for each test case for each participant. Then will rank teams for each cases separately. At the end final ranking will be the average of the ranking for each case.

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

Teams with missing submissions will be removed from the final ranking.

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

Following discussions with our biostatistician, and also while considering transparency and fairness to the participants.

Segmentation and its associated uncertainty is heterogeneous across patient cases. The above mentioned ranking scheme will allow us to see for each case individually which methods are better, and at the end check on an average which method is performing better on all test cases.

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.

We will conduct further permutation testing, to determine statistical significance of the relative rankings between each pair of teams. Specifically, for each team we start with a list of observed subject-level Cumulative Ranks, i.e., the actual ranking described above. For each pair of teams, we repeatedly randomly permute (i.e., 100,000 times) the Cumulative Ranks for each subject. For each permutation, we calculated the difference in the FRS between this pair of teams. The proportion of times the difference in FRS calculated using randomly permuted data exceeded the observed difference in FRS (i.e., using the actual data) indicated the statistical significance of their relative rankings as a p-value. These values will be reported in an upper triangular matrix.

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.

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.

B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, 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

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

S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, 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 (2018)

L. Maier-Hein, A. Reinke, M. Kozubek, A.L. Martel, T. Arbel, M. Eisenmann, et al., "BIAS: Transparent reporting of biomedical image analysis challenges", arXiv preprint arXiv:1910.04071 (2019)

Data Citations:

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection", The Cancer Imaging Archive, 2017. DOI: 10.7937/K9/TCIA.2017.KLXWJJ1Q

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection", The Cancer Imaging Archive, 2017. DOI: 10.7937/K9/TCIA.2017.GJQ7R0EF

K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, et al., "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository", Journal of Digital Imaging, 26(6):1045-1057 (2013)