HEad and neCK TumOR segmentation and outcome prediction in PET/CT images: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

HEad and neCK TumOR segmentation and outcome prediction in PET/CT images

Challenge acronym

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

HECKTOR (HEad and neCK TumOR segmentation)

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.

Head and Neck (H&N) cancers are among the most common cancers worldwide (5th leading cancer by incidence) (Parkin et al. 2005). Radiotherapy combined with cetuximab has been established as standard treatment (Bonner et al. 2010). However, locoregional failures remain a major challenge and occur in up to 40% of patients in the first two years after the treatment (Chajon et al. 2013). Recently, several radiomics studies based on Positron Emission Tomography (PET) and Computed Tomography (CT) imaging were proposed to better identify patients with a worse prognosis in a non-invasive fashion and by exploiting already available images such as these acquired for diagnosis and treatment planning (Vallières et al. 2017),(Bogowicz et al. 2017),(Castelli et al. 2017). Although highly promising, these methods were validated on 100-400 patients. Further validation on larger cohorts (e.g. 300-3000 patients) is required to ensure an adequate ratio between the number of variables and observations in order to avoid an overestimation of the generalization performance. Achieving such a validation requires the manual delineation of primary tumors and nodal metastases for every patient and in three dimensions, which is intractable and error-prone.
Methods for automated lesion segmentation in medical images were proposed in various contexts, often achieving expert-level performance (Heimann and Meinzer 2009), (Menze et al. 2015). Surprisingly few studies evaluated the performance of computerized automated segmentation of tumor lesions in PET and CT images (Song et al. 2013),(Blanc-Durand et al. 2018), (Moe et al. 2019).
In 2020, we organized the first HECKTOR challenge to offer an opportunity for participants working on 3D segmentation algorithms to develop automatic bi-modal approaches for the segmentation of H&N tumors in PET/CT scans, focusing on oropharyngeal cancers.
Following good participation and promising results in the 2020 challenge, we will increase the dataset size with 81 new cases provided by additional organization partners, from another clinical center with a different PET/CT scanner model and associated reconstruction settings (CHU Milétrie, Poitiers, France). In addition, we expand the scope of the challenge by considering an additional task with the purpose of outcome prediction based on the

PET/CT images. A clinically-relevant endpoint that can be leveraged for personalizing patient management at diagnosis will be considered: prediction of progression-free survival from diagnostic PET/CT images. By focusing on metabolic and morphological tissue properties respectively, PET and CT modalities include complementary and synergistic information for cancerous lesion segmentation as well as tumor characteristics relevant for patient outcome prediction, in addition to usual clinical variables (e.g., clinical stage, age, gender, treatment modality). Modern image analysis methods must be developed to best extract and leverage this information. The data used in this challenge is multi-centric, including four centers in Canada (Vallières et al. 2017), one center in Switzerland (Castelli et al. 2017), and one center in France (Hatt et al. 2019; Legot et al. 2018) for a total of 335 patients with annotated primary tumors.

Challenge keywords

List the primary keywords that characterize the challenge.

Head and Neck cancer; automatic segmentation; radiomics; PET/CT; multimodal

Year

The challenge will take place in ...

2021

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

None

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.

Based on last year's participation (18 participating teams) and the addition of a second task (outcome prediction), which will be of interest to other teams and may not require participating in the first task (segmentation), we can reasonably expect around 50 participants.

Publication and future plans

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

We plan to coordinate a publication (overview paper) describing the challenge results. A journal paper will be submitted to MedIA summarizing the results and outcomes of the challenge.
The leaderboard will remain open after the challenge for new submissions.

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.

The platform used for the online challenge will be AIcrowd (www.aicrowd.com/challenges/hecktor).
Standard equipment for presentations will be needed: projector, computer, loudspeakers and microphones.

TASK: Tumor segmentation

SUMMARY

Abstract

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

PET/CT head and neck primary tumor segmentation. This task is the same as in HECKTOR 2020, except on a larger, more diverse dataset.
The two tasks below are added to the 2021 edition.

Keywords

List the primary keywords that characterize the task.

Segmentation, PET/CT, head and neck cancer

ORGANIZATION

Organizers

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

- Vincent Andrearczyk (Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland)
- Valentin Oreiller (Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland AND Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland)
- Martin Vallières (Medical Physics Unit, McGill University, Montréal, Québec, Canada)
- Mathieu Hatt (LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France)
- Catherine Cheze-Le Rest (Nuclear medicine department, CHU Poitiers, Poitiers, France and LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France)
- Dimitris Visvikis (LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France)
- Mario Jreige (Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland)
- Hesham Elhalawani (Cleveland Clinic Foundation, Department of Radiation Oncology, Cleveland, OH, USA)
- Sarah Boughdad (Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland)
- John O. Prior (Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland)
- Adrien Depeursinge (Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland AND Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland)

b) Provide information on the primary contact person.

Vincent Andrearczyk vincent.andrearczyk@hevs.ch

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

AIcrowd

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

2021 will be opened later (2020 website: https://www.aicrowd.com/challenges/miccai-2020-hecktor)

Participation policies

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

Fully automatic.

Additional points: Algorithms producing fully-automatic segmentation of the test cases will be assessed.

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 can use the training data in any manner they would like for training the models. The use of additional (public or not) data for training should be reported along with the methodology description. Participants using additional data will not be eligible for the prize. We will split the evaluation and report all results on the website and in the publications.

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.

1st rank Award: 500 euros, conditioned on a paper submission reporting the details of the methods

e) Define the policy for result announcement.

Examples:

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

The results and winner will be announced publicly.
Once participants submit their results on the test set to the challenge organizers via the challenge website, they will be considered fully vested in the challenge, so that their performance results (without identifying the participant unless permission is granted) will become part of any presentations, publications, or subsequent

analyses derived from the Challenge at the discretion of the organizers.

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 participating teams are encouraged to publish their results in the LNCS proceedings of the challenge (following the MICCAI proceedings timeline and subject to acceptance). Participants can submit their results elsewhere when citing the overview paper, and (if so) no embargo will be applied.

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.

Segmentation outputs will be submitted by the participating teams via AIcrowd. We will provide a link to the submission instructions, also available on AIcrowd.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

The participating teams will be allowed to submit multiple results (with a limit of 5 submissions) to evaluate their algorithms. Only the best run will be officially counted to compute the challenge results. The participants will not receive feedback before the submission deadline (except for errors). We will add an online evaluation on a subset of training cases to allow the participants to check their outputs for the segmentation task.

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

The release date of the training cases: June 1 2021
The release date of the test cases: Aug. 1 2021
The submission date(s): opens Sept. 1 2021 closes Sept. 10 2021
Associated workshop days: Sept. 27 2021 or Oct. 1 2021
The release date of the results: Sept. 15 2021

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 indicate the answers for the cases of the different centers (list of centers is provided in 21.c).
Montreal: CHUM, CHUS, HGJ, HMR data (training): The ethics approval was granted by the Research Ethics Committee of McGill University Health Center (Protocol Number: MM-JGH-CR15-50).
Lausanne: CHUV data (testing): The ethics approval was obtained from the Commission cantonale (VD) d'éthique de la recherche sur l'être humain (CER-VD) with protocol number: 2018-01513.
Poitiers: CHUP data (partly training and testing): The fully anonymized data originates from patients who consent to the use of their data for research purposes.

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 NC ND.

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 code to produce the results and ranking is available on our GitHub repository. Link to the code and documentation will be added to the AIcrowd platform.

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

Although the participating teams will be strongly encouraged to disclose or share their code, it will remain optional.

Conflicts of interest

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

No conflict of interest applies.
The challenge is partly funded by the Swiss National Science Foundation (SNSF, grant 205320_179069). Siemens Switzerland Healthineers sponsored last year's challenge and will be contacted again to do so this year.
Only the organizers will have access to the test case ground truth contours/outcomes.

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

Prognosis, Research, Diagnosis.

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.

Patients received for initial staging of oropharyngeal H&N cancer. The clinical goals are two-fold; the automatically segmented regions can be used for (i) treatment planning in radiotherapy, (ii) further investigations to predict clinical outcomes such as progression-free survival (see task 2) and tumor aggressiveness.

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

Patients with histologically proven oropharyngeal H&N cancer who underwent radiotherapy and/or chemotherapy treatment planning.

Imaging modality(ies)

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

FDG-PET/CT scans (both functional FDG PET and low-dose CT modalities acquired with the same multimodal scanner are considered).

Context information

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

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

The information on image data will include clinical center, scanner information, DICOM meta-data including acquisition parameters and reconstruction algorithms.

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

None.

Target entity(ies)

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

The data originates from FDG-PET and low-dose non-contrast-enhanced CT images (acquired with combined PET/CT scanners) of the H&N region.

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.

The algorithm target is the H&N primary Gross Tumor Volume (GTVt). In particular, oropharyngeal cancers are considered in this challenge.

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.

Additional points: To perform well in this task, the algorithms to be optimized must find highly accurate H&N tumor segmentation for PET/CT images. Since the problem is imbalanced, we will consider appropriate metrics, i.e. Dice score coefficient as defined below (item 26).

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 device(s) used to acquire the training and testing images are listed in the following for the different centers (list of centers is provided in 21.c).
Montreal data (training):
HGJ: A hybrid PET/CT scanner (Discovery ST, GE Healthcare).
CHUS: A hybrid PET/CT scanner (GeminiGXL 16, Philips).
HMR: A hybrid PET/CT scanner (Discovery STE, GE Healthcare).
CHUM: A hybrid PET/CT scanner (Discovery STE, GE Healthcare).
Lausanne data (testing)
CHUV: A hybrid PET/CT scanner (Discovery D690 TOF, GE Healthcare).
Poitiers data (training and testing)
CHUP: A hybrid PET/CT scanner (Biograph mCT 40 ToF, Siemens).

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

HGJ:
For the PET portion of the FDG-PET/CT scan, a median of 584 MBq (range: 368-715) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 180-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a span (axial mash) of 5. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52 × 3.52 mm 2 (range: 3.52-4.69). For the CT portion of the FDG-PET/CT scan, an energy of 140 kVp with an exposure of 12 mAs was used. The CT slice thickness resolution was 3.75 mm and the median in-plane resolution was 0.98 × 0.98 mm 2 for all patients.
CHUS:
For the PET portion of the FDG-PET/CT scan, a median of 325 MBq (range: 165-517) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 150 s (range: 120-151) per bed position. Attenuation corrected images were reconstructed using a LOR-RAMLA iterative algorithm. The FDG-PET slice thickness resolution was 4 mm and the median in-plane resolution was 4×4 mm 2 for all patients. For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 12-140) with a median exposure of 210 mAs (range: 43-250) was used. The median CT slice thickness resolution was 3 mm (range: 2-5) and the median in-plane resolution was 1.17 × 1.17 mm 2 (range: 0.68-1.17).
HMR:
For the PET portion of the FDG-PET/CT scan, a median of 475 MBq (range: 227-859) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 360 s (range: 120-360) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 5 (range: 3-5). The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52 × 3.52 mm 2 (range: 3.52-

5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 120-140) with a median exposure of 11 mAs (range: 5-16) was used. The CT slice thickness resolution was 3.75 mm for all patients and the median in-plane resolution was 0.98 × 0.98 mm 2 (range: 0.98-1.37).
CHUM:
For the PET portion of the FDG-PET/CT scan, a median of 315 MBq (range: 199-3182) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 120-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 3 (range: 3-5). The median FDG-PET slice thickness resolution was 4 mm (range: 3.27-4) and the median in-plane resolution was 4 × 4 mm 2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 120 kVp (range: 120-140) with a median exposure of 350 mAs (range: 5-350) was used. The median CT slice thickness resolution was 1.5 mm (range: 1.5-3.75) and the median in-plane resolution was 0.98 × 0.98 mm 2 (range: 0.98-1.37).
CHUV:
The patients fasted at least 4h before the injection of 4 Mbq/kg of(18F)-FDG (Flucis). Blood glucose levels were checked before the injection of (18F)-FDG. If not contra-indicated, intravenous contrast agents were administered before CT scanning. After a 60-min uptake period of rest, patients were imaged with the PET/CT imaging system. First, a CT (120 kV, 80 mA, 0.8-s rotation time, slice thickness 3.75 mm) was performed from the base of the skull to the mid-thigh. PET scanning was performed immediately after acquisition of the CT. Images were acquired from the base of the skull to the mid-thigh (3 min/bed position). PET images were reconstructed by using an ordered-subset expectation maximization iterative reconstruction (OSEM) (two iterations, 28 subsets) and an iterative fully 3D (DiscoveryST). CT data were used for attenuation calculation.
CHUP:
PET/CT acquisition began after 6 hours of fasting and 60±5 min after injection of 3 MBq/kg of 18F-FDG (421±98 MBq, range 220-695 MBq). Non-contrast-enhanced, non-respiratory gated (free breathing) CT images were acquired for attenuation correction (120 kVp, Care Dose® current modulation system) with an in-plane resolution of 0.853×0.853 mm2 and a 5 mm slice thickness. PET data were acquired using 2.5 minutes per bed position routine protocol and images were reconstructed using a CT-based attenuation correction and the OSEM-TrueX-TOF algorithm (with time-of-flight and spatial resolution modeling, 3 iterations and 21 subsets, 5 mm 3D Gaussian post-filtering, voxel size 4×4×4 mm3).

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

HGJ: Hôpital général juif, Montréal, CA
CHUS: Centre hospitalier universitaire de Sherbooke, Sherbrooke, CA
HMR: Hôpital Maisonneuve-Rosemont, Montréal, CA
CHUM: Centre hospitalier de l'Université de Montréal, Montréal, CA
CHUV: Centre Hospitalier Universitaire Vaudois, CH
CHUP: Centre Hospitalier Universitaire de Poitiers, FR

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).

None.

Training and test case characteristics

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

Examples:

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

Training and testing cases represent one 3D FDG-PET volume registered with a 3D CT volume of the head and neck region, as well as a binary contour with the annotated ground truth tumors (only available for training cases to the participating teams). The labels represent the primary Gross Tumor Volume (GTVt). Patient information including gender and age is also included with each case.

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

The total number of training cases is 229. No specific validation cases are provided and the training set can be split in any manner for cross-validation. The total number of test cases is 106. A total of 81 cases were added to the previous year's dataset (28 and 53 to the train and test set respectively).

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

The proportion in 2020 was chosen due to the public availability of the training data (201 cases) and the number of non-public data for the test set (53 cases). This also allowed us to have test cases acquired from a different center from the training ones, while providing enough test cases for comparing the algorithms. For this new edition, we are including new data from another sixth center. We split it between training and test data to evaluate the prediction performance when the center was seen during training, a relevant clinical scenario too.
The split is performed to double the test size, i.e. 53 cases from CHUP are added to the test set, 28 to the training set.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Training and test cohorts are representative of the distribution of the real-world population of patients accepted for initial staging of oropharyngeal cancer (with ~25% of recurrence events and a median progression-free survival of ~30.7 months). The training/test split of the new center (CHUP) is performed with respect to the second and third tasks for stratified patient outcomes.

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.

Original annotations were performed differently depending on the centers:

Training set CHUV, CHUS, HGJ, HMR: Contours defining the GTVt were drawn by an expert radiation oncologist in a radiotherapy treatment planning system. 40\% (80 cases) of the training radiotherapy contours were directly drawn on the CT of the PET/CT scan and thereafter used for treatment planning. The remaining 60\% (121) of the training radiotherapy contours were drawn on a different CT scan dedicated to treatment planning and were then registered to the FDG-PET/CT scan reference frame using intensity-based free-form deformable registration with the software MIM (MIM software Inc., Cleveland, OH). For the training cases the original number of annotators is unknown.

Test set CHUV: For each patient in the test set, the GTV and metastatic lymph nodes were manually drawn on each FDG-PET/CT by a single expert radiation oncologist. Note that the segmentation of lymph nodes was performed with a separate delineation therefore the delineation of primary tumor only is available for all cases and is the only one exploited in the present challenge.

Training/test CHUP: the metabolic volume of tumors was automatically determined with the PET segmentation algorithm Fuzzy Locally Adaptive Bayesian (FLAB) (Hatt et al. 2009) and was then edited and corrected manually by a single expert based on the CT image, for example to correct cases where the PET-defined delineation included air or non-tumoral tissues in the corresponding CT.

IMPORTANT: A quality control was then performed by a single expert (certified as both a radiologist and a nuclear medicine physician) on all the datasets (training and testing) to ensure consistency in ground-truth contours definition. The expert re-annotated them, when necessary, to the real tumoral volume (often smaller than volumes delineated for radiotherapy).

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

The expert radiation oncologists annotated the images for treatment planning of radiotherapy, without specific instruction beyond the official contouring guidelines for radiotherapy planning. For the quality control, the annotator (who is both a radiologist and a nuclear medicine physician) was instructed to refine the original annotation to focus on radiologically suspicious cancer regions. Two non-experts (organizers of the challenge) performed an initial cleaning in order to facilitate the expert's work. The expert either validated or edited the VOIs.

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

The original annotators of the training and test cases were expert radiation oncologists with unknown expertise in terms of number of years of professional expertise. The annotator who performed the quality control is both an expert radiologist and nuclear medicine physician.

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.

A single annotation was performed, no merging was necessary.

Data pre-processing method(s)

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

The preprocessing involves (for both the training and test cases): (i) Computation of the Standardized Uptake Value for the PET images (ii) Conversion of the DICOM file format to NIfTI format (iii) Detection of the oropharyngeal region as a bounding box of 144x144x144 mm3. This region is automatically detected by locating the brain on the PET image and manually corrected when necessary. It is provided to the participants as a csv file for both the training and test set. The implementation details are provided in (Andrearczyk et al. 2020) and the code is available on the GitHub repository: https://github.com/voreille/hecktor.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

We evaluated the inter-observer agreement of 4 annotators including radiologist(s), nuclear medicine physician(s) and radiation oncologist(s) on a subset of 21 cases randomly drawn from the test and train data of HECKTOR 2020. The average DSC computed on the six pairs of annotators was 0.61. In the literature, (Gudi et al. 2017) reported similar agreement with an average inter-observer DSC of 57% and 69% on CT and PET/CT respectively for GTV segmentation (primary and lymph nodes). A source of error therefore originates from the degree of subjectivity in the annotation and correction of the expert.
For most patients of the dataset of HECKTOR 2020, the tumors were contoured on another CT scan, then the two CTs were registered (as described in 23a) and the annotations were transformed according to the registrations. Thus, a main source of error came from the registration on the original annotation. The quality control that we ran largely reduced this source of error (see item 23a). A similar quality control will be done on the new data, in order to homogenize the data.
Another source of error comes from the lack of CT images with a contrast agent for a more accurate delineation of the primary tumor.

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

None.

ASSESSMENT METHODS

Metric(s)

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

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

The Dice Similarity Coefficient (DSC) and Hausdorf distance (95%) will be performed on the 3D volumes to assess the segmentation algorithms by comparing the automatic segmentation and the annotated ground truth within the provided bounding boxes (see item 24). The final ranking will be based on the mean ranking across the test cases of the two metrics.
Precision and recall will also be computed to assess over- and under-segmentation, as well as the arithmetic mean

of sensitivity and positive predictive value.

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

DSC measures volumetric overlap between segmentation results and annotations. It is a good measure of segmentation for imbalanced segmentation problems, i.e. the region to segment is small as compared to the image size. DSC is commonly used in the evaluation of segmentation algorithms and particularly tumor segmentation tasks (Gudi et al. 2017), (Song et al. 2013), (Blanc-Durand et al. 2018), (Moe et al. 2019), (Menze et al. 2015). As mentioned in the challenge abstract, one aim of the developed algorithms is to further perform radiomics studies to predict clinical outcomes. DSC mostly evaluates the segmentation inside the ground truth volume (similar to intersection over union) and less the segmentation precision at the boundary. Therefore, DSC is particularly relevant for H&N radiomics where first and second order statistics are most relevant and less sensitive to small changes of the contour boundaries (Depeursinge et al. 2015). Shape features are expected to be less useful in H&N because the tumors are not spiculated and constrained by the anatomy of the throat.

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.

The ranking will be based on the average ranking of DSCs across all test cases. The method with the highest average DSC will be best. In case of equal DSC, we will consider the standard deviation around the average (smaller standard deviation wins).

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

In the unlikely event of missing results on the test case, DSCs of zero will be used for the corresponding missing results to compute the average score.

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

DSC is a commonly used metric for assessing automatic segmentation (Menze et al. 2015) and its preference over other metrics is further discussed in item 26b.

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.

A paired Wilcoxon signed-rank test will be used on the DSC per patient for statistical comparison of algorithms results. In the unlikely event of missing data (missing test segmentation results), DSCs of zero will be used for the corresponding missing results to compute the statistics. The python SciPy library will be used for the statistical analyses. For the multiple comparison testing (comparison of one vs multiple groups), correction will be used (Bonferroni). This test is computed to assess a significant difference between the first ranked algorithms and the following ones (2nd, 3rd etc.).

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

We will use a pair test since we want to compare the performance of the algorithms for each patient individually. The paired Wilcoxon signed-rank allows us to perform the analysis with minimal hypotheses.

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.

Similarly to last year, an ensemble using STAPLE (Warfield, Zou, and Wells 2004) will be performed with all the predictions from the participants, as well as only the best ones.
The correlation of the performance with the tumor size will be evaluated. The results will be compared across the two centers of the test set, one seen during training, the other one not, to evaluate the impact. In addition, an analysis of the location of segmentation false positives will be used to reveal if they are located in the vicinity of the primary tumor or in other regions of the oropharynx.

TASK: Radiomics

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.

Prediction of progression-free survival. The available data will be: i) PET and CT images similar to task 1 (acquired from multimodality integrated PET/CT devices) as well as ii) clinical variables.

Keywords

List the primary keywords that characterize the task.

Outcome prediction, radiomics, PET/CT, head and neck cancer

ORGANIZATION

Organizers

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

Same as task 1

b) Provide information on the primary contact person.

Mathieu Hatt hatt@univ-brest.fr

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

AIcrowd

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

Same as task 1

Participation policies

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

Fully automatic.

Additional points: Algorithms producing fully-automated prediction of progression-free survival of the test cases will be assessed.

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 can use the training data in any manner they would like for training the models. The use of additional (public or not) data for training should be reported along with the methodology description. Participants using additional data will not be eligible for the prize. We will split the evaluation and report all results on the website and in the publications.

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.

1st rank Award: 500 euros, conditioned on a paper submission reporting the details of the methods

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.

Same as task 1.

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).

Same as task 1.

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.

Prediction outputs will be submitted by the participating teams via AIcrowd. We will provide a link to the submission instructions, also available on AIcrowd.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

The participating teams will be allowed to submit multiple results (with a limit of 5 submissions) to evaluate their algorithms. Only the best run will be officially counted to compute the challenge results. The participants will not receive feedback before the submission deadline (except for errors).

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

Same as task 1.

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).

Same as task 1.

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 NC ND.

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.

same as task 1.

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

Same as task 1.

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.

Same as task 1.

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

Prognosis, Research, Diagnosis.

Task category(ies)

State the task category(ies).

Examples:

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

Prediction.

Cohorts

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

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

Patients received for initial staging of oropharyngeal H&N cancer. The clinical goal is the stratification of the population for personalized patient management.

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

Same as task 1.

Imaging modality(ies)

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

Same as task 1.

Context information

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

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

Same as task 1.

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

The patient information will include age, gender, treatment modality (chemotherapy in addition to radiotherapy or radiotherapy alone) and clinical stage.

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.

Same as task 1, plus clinical information.

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.

Same as task 1.

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.

Additional points: To perform well in this task, the algorithms to be optimized must correctly provide a reliable ranking of the survival times based on the individual risk scores.

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).

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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).

None.

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.

Same as task 1. In addition, the cases also include the patient outcome information (only available for training cases to the participating teams) of progression-free survival (time-to-event in days and censoring).

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

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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.

The outcome ground truths for the prediction task were collected in patients folders as registered by clinicians during patient follow-ups.

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.

No specific instructions. The patients' outcomes were collected as recorded in the patients folders at the last follow up.

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.

Not applicable.

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.

Not applicable.

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.

Same as task 1 for the images. No preprocessing of the patient outcomes is required.

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.

A source of error inherent to the task of survival analysis is the censored data (e.g. due to death or cessation of follow-up).
The median follow-up period is approximately 40 months (5-110 months).

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

None.

ASSESSMENT METHODS

Metric(s)

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

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

Concordance index (C-index).

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

The concordance index (C-index) quantifies the model's ability to provide an accurate ranking of the survival times based on the computed individual risk scores, generalizing the area under the ROC curve (AUC). It can account for censored data and represents the global assessment of the model discrimination power (Uno, et al. 2011).

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.

The ranking will be based on the C-index value obtained in the test cohort. The participant with the highest C-index wins.

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

In the unlikely event of a missing score for one patient, all the pairs containing the missing score will be treated as non-concordant.

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

The C-index is a widely used metric for survival analysis which allows taking into account time-to-event information and censored data. (Harrell 1982) (Uno et al. 2011)

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.

A bootstrap will be performed on the test set to evaluate the variance of the C-index of each team and for statistical comparison of algorithms results. The python SciPy and Scikit-learn libraries will be used for the statistical analyses.

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

The bootstrap analysis allows us to simulate different sampling of the test set to approximate the variance of the C-index.

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.

A consensus of the best models will be evaluated using the C-index, in order to establish whether there is complementary value for prognosis. Similarly to task 1, we will compare the results obtained on the two different centers of the test set (one center only seen during training).

TASK: Radiomics with ground truth contours

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.

Prediction of progression-free survival. Same as task 2 above, except ground-truth of primary tumors delineation will be provided in addition to images and clinical variables (i.e., participants only have to extract any features they wish from delineated primary tumors in order to build predictive models).

Keywords

List the primary keywords that characterize the task.

Same as task 2

ORGANIZATION

Organizers

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

Same as task 1

b) Provide information on the primary contact person.

Mathieu Hatt hatt@univ-brest.fr

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

AIcrowd

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

Same as task 1

Participation policies

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

Fully automatic.

Additional points: Same as task 2.

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 can use the training data in any manner they would like for training the models. The use of additional (public or not) data for training should be reported along with the methodology description. Participants using additional data will not be eligible for the prize. We will split the evaluation and report all results on the website and in the publications.

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.

1st rank Award: 500 euros, conditioned on a paper submission reporting the details of the methods

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.

Same as task 1.

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).

Same as task 1.

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.

Docker containers will be submitted by the participating teams via AIcrowd in order to generate the outcome prediction outputs. This is to allow the participants to provide solutions relying on ground-truth contours of the test data without actually providing these contours to participants. We will provide a link to the submission

instructions, also available on Alcrowd, and examples of containers to help the participants.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

The participating teams will be allowed to submit multiple results (with a limit of 5 submissions) to evaluate their algorithms. Only the best run will be officially counted to compute the challenge results. The participants will not receive feedback before the submission deadline (except for errors).

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

Same as task 1.

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).

Same as task 1.

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 NC ND.

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.

same as task 1.

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

Same as task 1.

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.

Same as task 1.

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

Prognosis, Research, Diagnosis.

Task category(ies)

State the task category(ies).

Examples:

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

Prediction.

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.

Same as task 2.

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

Same as task 1.

Imaging modality(ies)

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

Same as task 1.

Context information

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

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

Same as task 1.

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

Same as task 2.

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.

Same as task 2.

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.

Same as task 1.

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.

Additional points: Same as task 2.

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).

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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).

None.

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.

Same as task 2. In addition, ground-truth contours are provided for the algorithm but only through a docker framework to ensure the challengers do not have direct access to them.

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

Same as task 1.

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

Same as task 1.

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.

Same as task 1.

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.

Same as task 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.

Same as task 2 for the outcomes, same as task 1 for the tumor delineation.

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.

Same as task 1 for the tumor delineation

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.

Same as task 1 for the tumor delineation.

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.

Same as task 1 for the images and contours. No preprocessing of the patient outcomes is required.

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.

Same as tasks 1 and 2.

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

None.

ASSESSMENT METHODS

Metric(s)

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

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

Concordance index (C-index).

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

Same as task 2.

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.

Same as task 2.

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

Same as task 2.

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

Same as task 2.

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.

Same as task 2.

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

Same as task 2.

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.

Same as task 2. We will also compare the performance of Tasks 2 and 3, i.e. the influence of using ground truth tumor contours for outcome prediction or not.

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.

Andrearczyk, Vincent, Valentin Oreiller, and Adrien Depeursinge. "Oropharynx detection in PET-CT for tumor segmentation." Irish Machine Vision and Image Processing (2020).
Blanc-Durand, Paul, Axel Van Der Gucht, Niklaus Schaefer, Emmanuel Itti, and John O. Prior. 2018. "Automatic Lesion Detection and Segmentation of 18F-FET PET in Gliomas: A Full 3D U-Net Convolutional Neural Network Study." PloS One 13 (4): e0195798.
Bogowicz, Marta, Oliver Riesterer, Luisa Sabrina Stark, Gabriela Studer, Jan Unkelbach, Matthias Guckenberger, and Stephanie Tanadini-Lang. 2017. "Comparison of PET and CT Radiomics for Prediction of Local Tumor Control in Head and Neck Squamous Cell Carcinoma." Acta Oncologica 56 (11): 1531–36.
Bonner, James A., Paul M. Harari, Jordi Giralt, Roger B. Cohen, Christopher U. Jones, Ranjan K. Sur, David Raben, et al. 2010. "Radiotherapy plus Cetuximab for Locoregionally Advanced Head and Neck Cancer: 5-Year Survival Data from a Phase 3 Randomised Trial, and Relation between Cetuximab-Induced Rash and Survival." The Lancet Oncology 11 (1): 21–28.
Castelli, J., A. Depeursinge, V. Ndoh, J. O. Prior, M. Ozsahin, A. Devillers, H. Bouchaab, et al. 2017. "A PET-Based Nomogram for Oropharyngeal Cancers." European Journal of Cancer 75 (April): 222–30.
Chajon, Enrique, Caroline Lafond, Guillaume Louvel, Joël Castelli, Danièle Williaume, Olivier Henry, Franck Jégoux, et al. 2013. "Salivary Gland-Sparing Other than Parotid-Sparing in Definitive Head-and-Neck Intensity-Modulated Radiotherapy Does Not Seem to Jeopardize Local Control." Radiation Oncology. https://doi.org/10.1186/1748-717x-8-132.
Harrell, Frank E. 1982. "Evaluating the Yield of Medical Tests." JAMA: The Journal of the American Medical Association. https://doi.org/10.1001/jama.1982.03320430047030.
Hatt, Mathieu, Catherine Cheze le Rest, Alexandre Turzo, Christian Roux, and Dimitris Visvikis. 2009. "A Fuzzy Locally Adaptive Bayesian Segmentation Approach for Volume Determination in PET." IEEE Transactions on Medical Imaging 28 (6): 881–93.
Hatt, Mathieu, Florent Tixier, Marie-Charlotte Desseroit, Bogdan Badic, Baptiste Laurent, Dimitris Visvikis, and Catherine Cheze Le Rest. 2019. "Revisiting the Identification of Tumor Sub-Volumes Predictive of Residual Uptake after (chemo)radiotherapy: Influence of Segmentation Methods on F-FDG PET/CT Images." Scientific Reports 9 (1): 14925.
Heimann, Tobias, and Hans-Peter Meinzer. 2009. "Statistical Shape Models for 3D Medical Image Segmentation: A Review." Medical Image Analysis. https://doi.org/10.1016/j.media.2009.05.004.
Legot, Floriane, Florent Tixier, Minea Hadzic, Thomas Pinto-Leite, Christelle Gallais, Rémy Perdrisot, Xavier Dufour, and Catherine Cheze-Le-Rest. 2018. "Use of Baseline 18F-FDG PET Scan to Identify Initial Sub-Volumes with Local Failure after Concomitant Radio-Chemotherapy in Head and Neck Cancer." Oncotarget 9 (31): 21811–19.
Menze, Bjoern H., Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, et al. 2015. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)." IEEE Transactions on Medical Imaging 34 (10): 1993–2024.
Parkin, D. M., F. Bray, J. Ferlay, and P. Pisani. 2005. "Global Cancer Statistics, 2002." CA: A Cancer Journal for Clinicians. https://doi.org/10.3322/canjclin.55.2.74.
Song, Qi, Junjie Bai, Dongfeng Han, Sudershan Bhatia, Wenqing Sun, William Rockey, John E. Bayouth, John M. Buatti, and Xiaodong Wu. 2013. "Optimal Co-Segmentation of Tumor in PET-CT Images with Context Information." IEEE Transactions on Medical Imaging 32 (9): 1685–97.

Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. 2011. "On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data." Statistics in Medicine 30 (10): 1105–17.
Vallières, Martin, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Hugo J. W. L. Aerts, Nader Khaouam, et al. 2017. "Radiomics Strategies for Risk Assessment of Tumour Failure in Head-and-Neck Cancer." Scientific Reports 7 (1): 10117.
Warfield, Simon K., Kelly H. Zou, and William M. Wells. 2004. "Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation." IEEE Transactions on Medical Imaging 23 (7): 903–21.
Grossberg A, Mohamed A, Elhalawani H, Bennett W, Smith K, Nolan T, Chamchod S, Kantor M, Browne T, Hutcheson K, Gunn G, Garden A, Frank S, Rosenthal D, Freymann J, Fuller C.(2017). Data from Head and Neck Cancer CT Atlas. The Cancer Imaging Archive. DOI: 10.7937/K9/TCIA.2017.umz8dv6s
Moe, Yngve Mardal, Aurora Rosvoll Groendahl, Martine Mulstad, Oliver Tomic, Ulf Indahl, Einar Dale, Eirik Malinen, and Cecilia Marie Futsaether. "Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers." (2019).
Wee, L., & Dekker, A. (2019). Data from Head-Neck-Radiomics-HN1 [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.8kap372n.