3D Head and Neck Tumor Segmentation in PET/CT: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

3D Head and Neck Tumor Segmentation in PET/CT

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 reusing images 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 respect 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).

Therefore, it is timely to propose a MICCAI challenge to advance the methodological aspects and their validation for automated tumor and metastatic lymph nodes segmentation in PET/CT images. We also expect these progress and knowledge to be transferable for the segmentation of other types of cancer in the aforementioned imaging modalities. By focusing on metabolic and morphological tissue properties respectively, PET and CT modalities include complementary and synergistic information for cancerous lesion segmentation, that only modern image analysis methods can fully leverage.

This challenge will 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. Various approaches must be explored and compared to extract and merge information from the two

modalities, including early or late fusion, full volume or patch-based approaches, 2-, 2.5- or 3-D approach. The data used in this challenge will be multi-centric, including four centers in Canada (Vallières et al. 2017) and one center in Switzerland (Castelli et al. 2017) for a total of 249 patients with both tumor and metastatic lymph nodes contoured.
In addition to these 249 cases for which we already have all the agreements and information, we will likely include approximately 330 additional cases, including 215 public cases from (Grossberg et al. 2017), 88 public cases from (Wee et al. 2019) both for the training set, as well as approximately 30 non-public cases from McGill, for which we still need to obtain the data agreement. We do not include these cases in the data description as we do not have all data description or data agreement yet.

Challenge keywords

List the primary keywords that characterize the challenge.

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

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.

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 previous MICCAI challenges and general interest in 3D segmentation in the medical imaging community, we estimate the number of participants to be around 30.

Publication and future plans

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

An overview paper will be written with the organizing team's members. Each participating team who presented their method at the challenge session is allowed two co-authorships.
The leaderboard will remain open after the challenge.

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 online challenge will be AIcrowd (www.aicrowd.com/challenges/hecktor).
Standard equipment for presentations will be needed: projector, computer, loud speakers and microphones.

TASK: PET/CT head and neck 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.

Since we have a single task, please refer to the main abstract that fully covers this task.

Keywords

List the primary keywords that characterize the task.

Head and neck cancer; automatic segmentation; PET/CT; multimodal

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)
- Joel Castelli (Radiotherapy Department, Cancer Institute Eugène Marquis, 35000 Rennes, France AND INSERM, U1099, 35000 Rennes, France AND University of Rennes 1, LTSI, 35000 Rennes, France)
- Hesham Elhalawani (Cleveland Clinic, Cleveland, Ohio, USA)
- Mario Jreige (Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland)
- 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.

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.

aicrowd.com

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

www.aicrowd.com/challenges/hecktor

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.

Publicly available data is allowed.

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

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.

The 1st ranked team will receive an award of 500 euros.

e) Define the policy for result announcement.

Examples:

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

The top three performing teams 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).

An overview paper will be written with the organizing team's members. Each participating team who presented their method at the challenge session is allowed two co-authorships. The participating teams are encouraged to publish their results separately elsewhere (e.g. CEUR) 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 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 last run will be officially counted to compute the challenge results.

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: 01/06/2020
The release date of the test cases: 01/08/2020
The submission date(s): opens 1/09/2020 closes 10/09/2020
Associated workshop days: 4/10/2020 or 8/10/2020
The release date of the results: 15/09/2020

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).

Training dataset: The ethics approval was granted by the Research Ethics Committee of McGill University Health Center (Protocol Number: MM-JGH-CR15-50).
Test dataset: 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.

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 will be 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.

The participating teams will decide whether they want to disclose their code.

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).
Only the organizers will have access to the test case ground truth contours.

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, Diagnosis, Prognosis.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

Patients with suspected H&N cancer. The clinical goals are two-fold; the automatically segmented regions can be used for (i) treatment planning in radiotherapy, (ii) further radiomics studies to predict clinical outcomes such as overall patient survival, disease free survival, tumor aggressivity.

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

Patients with histologically proven H&N cancer who underwent radiotherapy treatment planning.

Imaging modality(ies)

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

FDG-PET/CT scans.

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).

The patient information will include age and gender.

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 CT images 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 participating algorithms will be designed to segment H&N tumor volumes, i.e. the union of Gross Tumor Volume (GTV) and metastatic lymph nodes volumes. In particular, oropharyngeal cancer 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 the challenge, 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 listed 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 test data are listed in the following for the different centers (the list of centers is provided in 21.c).
Training:
The devices used in the training dataset for the four cohorts are listed in the following.
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).

Test:
CHUV: A hybrid PET/CT scanner (Discovery D690 TOF, GE Healthcare).

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. 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. 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. 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. 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 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.

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

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 test 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 will represent the union of the GTV and metastatic lymph nodes. 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 203. 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 46.
In addition to the 203 training cases for which we already have all the agreements and information, we will likely include approximately 330 additional cases, including 215 public cases from (Grossberg et al. 2017) and 88 public cases from (Wee et al. 2019).
Similarly, in addition to the 46 test cases, we will likely include approximately 30 non-public cases from McGill University, for which we still need to obtain the ethics agreement.

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

This proportion was chosen due to the public availability of the training cases (the test cases are not), the possibility of testing on a different center and with a sufficiently large number of cases (46). This represents about 80% of the data for training and 20% for the test.

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.

None.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Training set: Contours defining the GTV and metastatic lymph nodes were drawn by an expert radiation oncologist in a radiotherapy treatment planning system. For 65 of the 203 patients, the radiotherapy contours were directly drawn on the CT scan of the FDG-PET/CT scan. For 138 of the 203 patients, the radiotherapy contours were drawn on a different CT scan dedicated to treatment planning. In the latter case, the contours were propagated to the FDG-PET/CT scan reference frame using deformable registration with the software MIMR©(MIM software Inc., Cleveland, OH).
For the training cases, we do not know the original number of annotators. However, we are currently organizing a quality control of the entire dataset (training and test) with a single annotator who is both radiologist and nuclear medicine physician.
Test set: 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.

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. For the quality control, the annotator (who is both radiologist and nuclear medicine physician) will be instructed to refine the original annotation to focus on radiologically suspicious cancer regions.

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. However, we are currently organizing a quality control of the entire dataset with a single annotator who is both 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:
(ii) Cropping of a 150x150x150mm volume that contains the GTV and metastatic lymph nodes centered around the oropharyngeal region. This region is automatically detected by locating the brain on the PET image. The implementation details will be provided in the GitHub repository.

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 will evaluate the inter-observer agreement of 4 annotators including radiologist(s), nuclear medicine physician(s) and radiation oncologist(s) on a subset of ~20/30 cases to assess human performance. According to (Gudi et al. 2017), we can expect an inter-observer DSC in GTV segmentation of 57% and 69 % on CT and PET/CT respectively.
For most patients, 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 comes from the registration on the original annotation. The quality control that we are currently running will largely reduce this source of error (see item 23a).

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) will be performed on the 3D volumes to assess the segmentation algorithms by comparing the automatic segmentation and the annotated ground truth. DSC will be used to compute the ranking.

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 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 DSC across all test cases. The method with highest average DSC will be best.

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.

T-tests and confidence intervals (CI) on average DSCs will be performed 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).

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

Parametric t-tests and CIs will be used as we compare averages and the central limit theorem (CLT) holds.

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 (Menze et al. 2015), we will evaluate simple ensembling methods; (i) the best combination: an upper limit of ensembling performance, selecting for each test case the best segmentation among all predictions. (ii) a hierarchical majority vote.

Inter-algorithm variability will be evaluated by comparing pairs of segmentations.

Common problems and biases of the submitted results will be assessed by averaging the scores across algorithms for each test case. The worst and best segmented images on average will be reported and visually inspected.

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.

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.

Depeursinge, Adrien, Masahiro Yanagawa, Ann N. Leung, and Daniel L. Rubin. 2015. “Predicting Adenocarcinoma Recurrence Using Computational Texture Models of Nodule Components in Lung CT.” Medical Physics. https://doi.org/10.1118/1.4916088.

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

Gudi, Shivakumar, Sarbani Ghosh-Laskar, Jai Prakash Agarwal, Suresh Chaudhari, Venkatesh Rangarajan, Siji Nojin Paul, Rituraj Upreti, Vedang Murthy, Ashwini Budrukkar, and Tejpal Gupta. 2017. “Interobserver Variability in the Delineation of Gross Tumour Volume and Specified Organs-at-Risk During IMRT for Head and Neck Cancers and the Impact of FDG-PET/CT on Such Variability at the Primary Site.” Journal of Medical Imaging and Radiation Sciences 48 (2): 184–92.

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.

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.

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).

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.

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.

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.