Head and neck tumor segmentation in PET/CT: The HECKTOR challenge

Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallieres, Simeng Zhu, Juanying Xie, Ying Peng, et al.

Medical Image Analysis 77 (2022) 102336

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Head and neck tumor segmentation in PET/CT: The HECKTOR challenge

Valentin Oreiller[a,b,*], Vincent Andrearczyk[a,1], Mario Jreige[b], Sarah Boughdad[b], Hesham Elhalawani[c], Joel Castelli[d], Martin Vallières[e], Simeng Zhu[f], Juanying Xie[g], Ying Peng[g], Andrei Iantsen[h], Mathieu Hatt[h], Yading Yuan[i], Jun Ma[j], Xiaoping Yang[k], Chinmay Rao[l], Suraj Pai[l], Kanchan Ghimire[m], Xue Feng[m,n], Mohamed A. Naser[o], Clifton D. Fuller[o], Fereshteh Yousefirizi[p], Arman Rahmim[p], Huai Chen[q], Lisheng Wang[q], John O. Prior[b], Adrien Depeursinge[a,b]

[a] *Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland*
[b] *Department of Nuclear Medicine and Molecular Imaging, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland*
[c] *Department of Radiation Oncology, Brigham and Women's Hospital and Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*
[d] *Radiotherapy Department, Cancer Institute Eugène Marquis, Rennes, France*
[e] *Department of Computer Science, Université de Sherbrooke, Sherbrooke, Québec, Canada*
[f] *Department of Radiation Oncology, Henry Ford Cancer Institute, Detroit, MI, USA*
[g] *School of Computer Science, Shaanxi Normal University, Xi'an 710119, PR China*
[h] *LaTIM, INSERM, UMR 1101, University Brest, Brest, France*
[i] *Department of Radiation Oncology, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
[j] *Department of Mathematics, Nanjing University of Science and Technology, Jiangsu, China*
[k] *Department of Mathematics, Nanjing University, Jiangsu, China*
[l] *Department of Radiation Oncology (Maastro), GROW School for Oncology, Maastricht University Medical Centre+, Maastricht, The Netherlands*
[m] *Carina Medical, Lexington, KY, 40513, USA*
[n] *Department of Biomedical Engineering, University of Virginia, Charlottesville VA 22903, USA*
[o] *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA*
[p] *Department of Integrative Oncology, BC Cancer Research Institute, Vancouver BC, Canada*
[q] *Department of Automation, Institute of Image Processing and Pattern Recognition, Shangai Jiao Tong University, Shanghai 200240, People's Republic of China*

ARTICLE INFO

Article history:
Received 22 April 2021
Revised 13 October 2021
Accepted 14 December 2021
Available online 25 December 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Medical imaging
Head and neck cancer
Oropharynx
Automatic segmentation
Challenge

ABSTRACT

This paper relates the post-analysis of the first edition of the HEad and neCK TumOR (HECKTOR) challenge. This challenge was held as a satellite event of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2020, and was the first of its kind focusing on lesion segmentation in combined FDG-PET and CT image modalities. The challenge's task is the automatic segmentation of the Gross Tumor Volume (GTV) of Head and Neck (H&N) oropharyngeal primary tumors in FDG-PET/CT images. To this end, the participants were given a training set of 201 cases from four different centers and their methods were tested on a held-out set of 53 cases from a fifth center. The methods were ranked according to the Dice Score Coefficient (DSC) averaged across all test cases. An additional inter-observer agreement study was organized to assess the difficulty of the task from a human perspective. 64 teams registered to the challenge, among which 10 provided a paper detailing their approach. The best method obtained an average DSC of 0.7591, showing a large improvement over our proposed baseline method and the inter-observer agreement, associated with DSCs of 0.6610 and 0.61, respectively. The automatic methods proved to successfully leverage the wealth of metabolic and structural properties of combined PET and CT modalities, significantly outperforming human inter-observer agree-

* Corresponding author at: Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland.
E-mail address: valentin.oreiller@hevs.ch (V. Oreiller).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.media.2021.102336

ment level, semi-automatic thresholding based on PET images as well as other single modality-based methods. This promising performance is one step forward towards large-scale radiomics studies in H&N cancer, obviating the need for error-prone and time-consuming manual delineation of GTVs.

1. Introduction

High-throughput medical image analysis, often referred to as radiomics, has shown its potential in unveiling relationships between quantitative image biomarkers and cancer prognosis, including in the context of Head and Neck (H&N) cancer (Vallieres et al., 2017; Bogowicz et al., 2017). H&N cancer is the 5th leading cancer by incidence (Parkin et al., 2005) and its treatment is generally based on a combination of radiotherapy with systemic treatment (e.g. Cetuximab) (Bonner et al., 2010). However, treating this cancer remains challenging since local failure occurs in about 40% of patients in the first two years after the treatment (Chajon et al., 2013). The development of non-invasive and personalized approaches (e.g. radiomics) is critical for improving disease characterization and will, hopefully, lead to more targeted therapies based on phenotypic tumor characteristics. 2-[18F]fluoro-2-deoxyglucose positron-emission tomography (FDG-PET) and Computed Tomography (CT) hold a special place for disease characterization since they contain complementary information about the metabolism and the anatomy of cancer. Furthermore, they are used for initial staging and follow-up of H&N cancer. These modalities are therefore readily available for the creation and evaluation of radiomics models based on these clinically acquired images. Typical radiomics analyses rely on localized feature extraction inside delineated lesions or Volumes Of Interest (VOI) (Lambin et al., 2017; Gillies et al., 2016). One of the reasons that impede the development of robust models is the time-consuming and error-prone manual delineation of these VOIs. To this end, the automatic segmentation of H&N Gross Tumor Volume of the primary tumor (GTVt) and the lymph nodes (GTVn) constitutes a highly promising approach to annotate and analyze very large cohorts, which is critically needed to enable robust and reproducible validation of radiomics models. Moreover, automatic segmentation also has the potential to allow radiation oncologists to improve treatment planning efficiency by reducing the time needed for tumor delineation as well as improving inter-observer reproducibility.

The goal of the HEad and neCK TumOR (HECKTOR) challenge is to establish and benchmark the best-performing methods for H&N lesions segmentation while exploiting the rich bi-modal information of combined PET/CT. In this first edition of the challenge, the participants were asked to develop automatic methods for the segmentation of the GTVt[2] on FDG-PET/CT images of patients suffering from oropharyngeal cancer. It is worth noting that to be part of the official ranking, the participants had to provide a paper describing their methods. Furthermore, participants had to disclose the use of external training data and were in this case not eligible for the official ranking. None of the participants reported using external data. This manuscript summarizes the methods and presents the associated segmentation results of the different teams who participated in this 2020 edition of the HECKTOR challenge. It also includes several additional extensive qualitative and quantitative analyses. This paper extends the material presented in (Andrearczyk et al., 2021b) with the following:

- an extensive review of the prior work;
- an analysis of the inter-observer agreement organized with four different observers on a subset of 21 cases;
- an evaluation of a super-ensemble segmentation based on the submitted contours of the ten ranked teams;
- an addition of new participants' results from runs submitted after the end of the challenge;
- a semi-automatic segmentation based on PET thresholding as an additional baseline; and
- additional extensive qualitative and quantitative analyses of the results.

The paper is organized as follows. Section 2 presents the related work. Section 3 describes the challenge setup including the dataset, annotations, participation, and ranking. The presentation and in-depth analysis of the participants' results are provided in Section 4 and are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Prior work

2.1. Related tumor segmentation algorithms

An abundance of works has been proposed to automatically segment tumors in PET and PET/CT images ranging from thresholding to unsupervised and supervised machine learning methods. Making an exhaustive review of all these approaches is out of the scope of this manuscript and is proposed in (Foster et al., 2014; Hatt et al., 2017). Among these different strategies, the simplest ones are based on the thresholding of the Standardized Uptake Values (SUV) in PET images. These methods are difficult to automatize completely since the SUV is a semi-quantitative measure that highly depends on the time between the injection and the image acquisition, the device, the reconstruction algorithm, the shape of the tumor, and even the patient (Wahl et al., 2009).

More refined approaches have been proposed to further automatize this process. Most of them are relying on the distribution of SUV values or other handcrafted quantitative image features in PET only. For instance, algorithms based on Gaussian Mixtures (Aristophanous et al., 2007) or fuzzy C-means modeling (Hatt et al., 2009; Lapuyade-Lahorgue et al., 2015) were proposed. Others formulated the segmentation problem as a minimization of a Markov random field (Song et al., 2013). In the context of H&N tumors delineation, a decision-tree-based K-nearest-neighbor classifier trained with regional texture features in PET and CT images was used in (Yu et al., 2009).

Recent work was inspired by the success of deep Convolutional Neural Networks (CNN), and more precisely of the U-Net (Ronneberger et al., 2015) applied to multi-modal biomedical image segmentation (Zhou et al., 2019). PET/CT tumor segmentation has also benefited from the advancement of this field. For instance, (Blanc-Durand et al., 2018) applied a 3D U-Net to segment brain tumors in O-(2-[18F]fluoroethyl)-L-tyrosine PET/CT images. Deep CNNs was also used several times in the context of lung tumor segmentation (Wu et al., 2020; Fu et al., 2021; Li et al., 2019; Zhao et al., 2018; Zhong et al., 2018). A 3D U-Net was used by (Jemaa et al., 2020) to lung cancer and lymphoma, which was

[2] For the first and second edition of the challenge, the GTVn segmentation is not part of the tasks but will be asked in further editions.

trained on 2540 volumes and tested 1124 volumes. (Iantsen et al., 2021a) used a U-Net architecture for the automatic segmentation of cervical tumors in PET only.

The deep learning-based approaches were also specifically applied to tumor segmentation in H&N cancers. A comparison of different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation is presented in (Ren et al., 2021). In a study including 22 patients from two different centers, (Huang et al., 2018) used a 2D U-Net to segment the GTV, *i.e.* the union of GTVt and GTVn. (Moe et al., 2019) used a 2D U-Net for the segmentation of GTV on a dataset of 55 patients. In another study, (Guo et al., 2019) applied a 3D U-Net to segment the GTVt, which was evaluated on a cohort of 250 patients. The authors showed that multimodal networks outperform networks based on a single modality. More recently, (Groendahl et al., 2021) performed an analysis of the different types of automatic segmentation based on thresholding, classification at the pixel level using a shallow classifier, and deep CNN methods. They did this comparison on a mono-centric cohort of 197 patients and concluded that deep learning models outperform the others.

Identifying the best performing method among all these different strategies requires a standardized evaluation. This was already highlighted by (Hatt et al., 2017) and challenges constitute a suitable way to systematically evaluate and compare state-of-the-art algorithms against the same test set and with highly controlled conditions.

2.2. Medical image segmentation challenges

The growing interest in biomedical image analysis challenges is illustrated by and an increasing number of new challenges organized every year, which can be partly explained by the growing community. For instance at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2018, 2019, and 2020 there were 15, 22, and 25 accepted challenges, respectively. In the past three MICCAI editions, 52 out of 125 tasks (42%) were related to segmentation.[3] Several other challenges are organized as satellite events of other conferences including the International Symposium on Biomedical Imaging (ISBI), the international conference on Medical Imaging with Deep Learning (MIDL), and the annual meeting of the Radiological Society of North America (RSNA), as well as independently organized challenges (e.g. on Kaggle[4]). Remarkably successful challenges in medical image segmentation include the Brain Tumor Segmentation (BraTS) challenge (Menze et al., 2014), Kidney Tumor Segmentation (KiTS) (Heller et al., 2021) challenge and the Visual Concept Extraction Challenge in Radiology (VISCERAL) (del Toro et al., 2014) challenge. Surprisingly, as of 2021, only one challenge was organized on PET segmentation (Hatt et al., 2018) and, to the best of our knowledge, none on PET/CT segmentation.

3. HECKTOR 2020 challenge set-Up

The challenge took place in 2020 and was associated with the 23rd MICCAI conference as a satellite event the same year. It was hosted on the AIcrowd platform.[5] The training and test data were released on the 10th of June and the 1st of August, respectively. The participants were asked to submit their results before the 10th of September. The challenge's results were communicated the 15th of September, and the MICCAI associated event was held the 4th of October. The data of the challenge are currently available on the AIcrowd platform after signing an end-user agreement and the leaderboard submission was open until the 10th of September 2021.[6]

Table 1
List of scanners used in the different centers.

Center	Device
HGJ	hybrid PET/CT scanner (Discovery ST, GE Healthcare)
CHUS	hybrid PET/CT scanner (GeminiGXL 16, Philips)
HMR	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUM	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUV	hybrid PET/CT scanner (Discovery D690 TOF, GE Healthcare)

The following section summarizes the challenge's set-up. A thorough and BIAS (Maier-Hein et al., 2020) compliant description of the challenge organization is provided in (Andrearczyk et al., 2021b).

3.1. Dataset

The dataset used in this challenge includes PET and CT images as well as patient information including age, sex, and acquisition center. The patients selected for this dataset suffered from H&N cancer, which was histologically proven, and they underwent radiotherapy treatment often combined with chemotherapy. The data were acquired from five centers:

1. Hôpital Général Juif (HGJ), Montréal, CA ($n = 55$)
2. Centre Hospitalier Universitaire de Sherbooke (CHUS), Sherbrooke, CA ($n = 72$)
3. Hôpital Maisonneuve-Rosemont (HMR), Montréal, CA ($n = 18$)
4. Centre Hospitalier de l'Université de Montréal (CHUM), Montréal ($n = 56$)
5. Centre Hospitalier Universitaire Vaudois (CHUV), CH ($n = 53$)

The four centers HGJ, CHUS, HMR, and CHUM were used for the training set, which amounts to 201 cases. This training data constitute a subset of (Vallieres et al., 2017) which contains 298 cases including H&N cancers originating from various anatomical regions. For this initial edition of the HECKTOR challenge, we decided to focus on patients suffering from oropharyngeal cancer to reduce anatomical variations and provide more controlled conditions for the algorithms. The CHUV center was used for the test set, totaling a number of 53 test cases.

An example of fused PET/CT images for each of the five centers is depicted in Fig. 1. The list of scanners used in each center for image acquisition can be found in Table 1. Additional information concerning image protocols are described in (Andrearczyk et al., 2021b).

The Digital Imaging and Communications in Medicine (DICOM) files were converted to the Neuroimaging Informatics Technology Initiative (NIfTI) format. The CT and PET images were stored in Hounsfield Units (HU) and SUVs, respectively. The code used for the conversion is available on the challenge's repository[7] Each case comprises NIfTI files for the CT image, the PET image, and the GTVt mask (for the training cases), as well as patient information (age, sex) and center. A bounding box locating the oropharyngeal region was also provided (details of the automatic region detection can be found in Andrearczyk et al., 2020a). The choice of preprocessing (*e.g.* resampling, image standardization) was left to the participants. Therefore, no further preprocessing was performed to mimic

[3] https://www.biomedical-challenges.org/miccai2021/Statistics, as of October 2021.
[4] https://www.kaggle.com/, as of October 2021.
[5] https://www.aicrowd.com/challenges/miccai-2020-hecktor, as of October 2021.
[6] The leaderboard was replaced by the 2021 edition after this date: https://www.aicrowd.com/challenges/miccai-2021-hecktor/leaderboards.
[7] github.com/voreille/hecktor/blob/hecktor2020/src/data/dicom_conversion.py, as of October 2021.

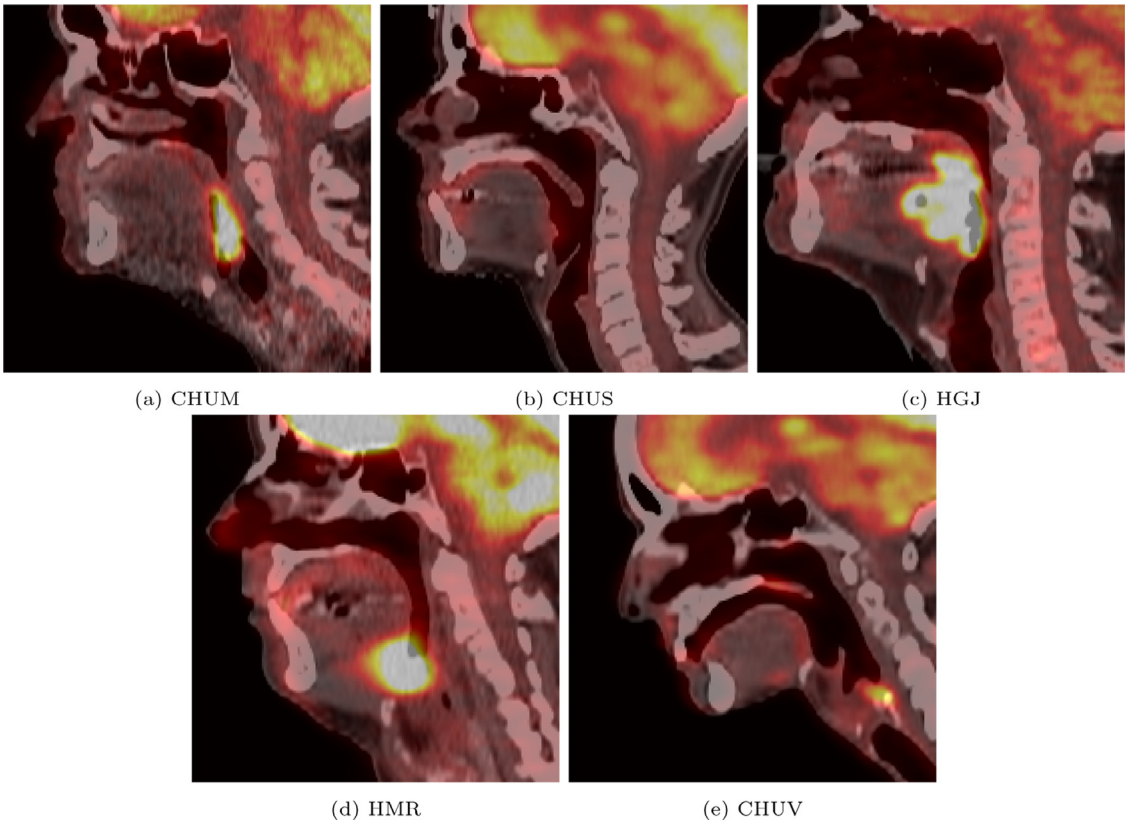


Fig. 1. Case examples of 2D sagittal slices of fused PET/CT images from each of the five centers. These images are obtained after resampling the PET image and the CT image to 1x1x1 mm^3 with a tricubic interpolation. The CT window in Hounsfield unit is $[-140, 260]$ and the PET window in SUV is $[0, 12]$.

a clinical use of the segmentation methods. However, we provided some routines to crop, resample, and also train a baseline CNN (using NiftyNet Gibson et al., 2018). This code was made available on the challenge's repository[8] to help the participants and to maximize transparency, but the participants were free to use their methods.

3.2. Contours

The GTVts from the original dataset were drawn by expert radiation oncologists from multiple centers for radiotherapy treatment planning. In most cases, the contours used for treatment planning are larger than the actual tumor and are presumably not optimized for radiomics with sometimes the inclusion of surrounding tissue or even air cavities. Furthermore, only 40% (80 cases) of the training set were delineated on the CT of the PET/CT scans. The remaining 60% were drawn on a dedicated CT scan for the treatment planning and were registered to the PET/CT scans using intensity-based free-form deformable registration with the software MIM (MIM Software Inc., Cleveland, OH). For more information about the original training set, please refer to (Vallieres et al., 2017). The original contours of the test set were all drawn on the fused PET/CT scans.

To homogenize the data *i.e.* to obtain delineations closer to the true tumoral volume and to remove variability due to the annotators and the registration step, each contour was controlled by an expert who is both a radiologist and a nuclear physician. Two non-experts annotators made an initial cleaning to facilitate the expert's work. During this control, multiple contours were rectified to follow the true border of the tumor as close as possible. Many original contours included air as well as various tissues around the tumor. In some cases, the registration between the dedicated CT planning and the PET/CT introduced artifacts that did not belong to the GTVt. In many cases, the GTVt and GTVn were stored under the same label and had to be separated. Three annotations were corrupted and could not be loaded, requiring the contours to be drawn from scratch. Among the 53 test cases, 11 images were contoured from scratch with the help of the radiological report.

Despite the high inter-observer variability (see Section 4.4), and with a slight misuse of language, we refer to these "controlled" reference annotations as ground truth.

Finally, the same VOI quality control process was performed for the GTVn contours. These contours were not directly used for the HECKTOR 2020 challenge but we used them in post-analysis of the results (see Section 4.8). We also plan on using these annotations in future editions as an auxiliary task of lymph node segmentation. Radiomics studies including lymph nodes may carry important information about patient prognosis and response to treatment.

3.3. Ranking and assement method

Participants were given access to the test cases without the ground truth annotations and were asked to submit the results of their algorithms on these cases on the AIcrowd platform. We only accepted binary segmentations in the NIfTI file format.

Results were ranked using the 3D Dice Similarity Coefficient (DSC) computed on images cropped using the provided bounding boxes (see Section 3.1) in the original CT resolution as:

$$DSC = \frac{2TP}{2TP + FP + FN}, \quad (1)$$

where TP, FP, and FN are the number of True Positive, False Positive, and False Negative at the voxel level, respectively. Prior to the

[8] github.com/voreille/hecktor/tree/hecktor2020, as of October 2021.

Table 2
Summary of the algorithms in terms of main components used: 2D or 3D U-Net, resampling, preprocessing, training or testing data augmentation, loss used for optimization, an ensemble of multiple models for test prediction and postprocessing of the results. We use the following abbreviations for the preprocessing: Clipping (C), Standardization (S), and if it is applied only to one modality, it is specified in parentheses. For the image resampling, we specify whether the algorithms use Isotropic (I) or Anisotropic (A) resampling and Nearest Neighbor (NN), Linear (L), or Cubic (Cu) interpolation. We use the following abbreviation for the losses: Cross-Entropy (CE), Mumford-Shah (MS), and Mean Absolute Error (MAE). More details can be found in the respective participants' publications.

Team	2D/3D	preproc.	resampling	augm.	loss	ensemble	postproc.
`andrei.iantsen` (Iantsen et al., 2021b)	3D	C+S	I/L	√	soft Dice+Focal	√	✗
`junma` (Ma and Yang, 2021)	3D	S(PET)	I/Cu	✗	Dice+Top-K	√	√
`badger` (Xie and Peng, 2021)	3D	C(CT)+S(PET)	A/Cu	√	Dice+CE	✗	✗
`deepX` (Yuan, 2021)	3D	C(CT)+S	I/L	√	Jaccard distance	√	✗
`AIView_sjtu` (Chen et al., 2021)	3D	C+S	A/NN	√	Dice	✗	✗
`xuefeng` (Ghimire et al., 2021)	3D	C(CT)+S	A/L	√	Dice+CE	√	√
`QuritLab` (Yousefirizi and Rahmim, 2021)	3D	S	I/L	✗	MS+MAE	✗	✗
`HFHSegTeam` (Zhu et al., 2021)	2D	C+S	I/L	√	soft Dice	✗	✗
`Fuller_MDA_Lab` (Naser et al., 2021)	3D	C+S	A/Cu	√	Dice+CE	✗	✗
`Maastro-Deep-Learning` (Rao et al., 2021)	2D/3D	C	A/Cu	✗	Top-K	√	√
`Our baseline 3D PET/CT` (Andrearczyk et al., 2020b)	3D	C+S	I/Cu	✗	Dice+CE	✗	✗
`Our baseline 2D PET/CT` (Andrearczyk et al., 2020b)	2D	C+S	I/Cu	✗	Dice+CE	✗	✗

challenge opening, we decided to handle missing predictions by attributing a DSC of 0 to them. However, this never happened during the submission phase. If the submitted results were in a resolution different from the CT resolution, we applied nearest-neighbor interpolation before evaluation. We also computed other metrics for comparison, namely precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$) to investigate whether the methods were rather providing a large FP or FN rate. The evaluation implementation can be found on our GitHub repository[9] and was provided to the participants to maximize transparency.

Each participating team had the opportunity to submit up to five valid runs, in case of formatting errors the participant was informed by an error message and the run was not counted. No immediate feedback was displayed on how their run was performing to avoid iterative overfit. The best result of each team was used in the final ranking, which is detailed in Section 4 and discussed in Section 5.

4. Results

This section regroups results in terms of challenge participation, algorithms used, segmentation performance, inter-observer agreement, ensembling "super-algorithm", simple PET thresholding, the relation between tumor size and segmentation performance, false-positive analysis, and alternative ranking of the methods.

4.1. Participation

The number of registered teams, as of September 10, 2020 (submission deadline), was 64. At the same date, we had also received and approved 85 signed end-user agreements, received 83 results submissions, including valid and invalid submissions. For the first iteration of the challenge, these numbers are high and show an important interest in the task.

4.2. Algorithms summary

Baselines We trained several baseline models using standard 3D and 2D U-Nets as in our preliminary results in (Andrearczyk et al., 2020b). It is worth noting that (Andrearczyk et al., 2020b) used a dataset that was different from HECKTOR 2020, and that the same algorithms were re-trained and evaluated using the HECKTOR 2020 data. We trained on multi-modal PET/CT as well as individual modalities with a combination of non-weighted Dice and cross-entropy losses and without data augmentation.

Participants' methods In Table 2, we summarize some of the main components of the participants' algorithms, including model architecture, preprocessing, training scheme and postprocessing. We only report the methods of the participants with an associated publication, which was crucial to ensure the scientific relevance of the challenge. More details on the individual methods can be found in Appendix A as well as in the corresponding participants' papers (Iantsen et al., 2021b; Chen et al., 2021; Ma and Yang, 2021; Rao et al., 2021; Xie and Peng, 2021; Zhu et al., 2021; Ghimire et al., 2021; Yousefirizi and Rahmim, 2021; Yuan, 2021; Naser et al., 2021). In the results Section 4, we also include results of the participants without publication for comparison.

All the participants used a U-Net-based architecture. Eight used 3D architectures, one used a 2D architecture and one used a combination of the two. All participants used some sort of preprocessing prior to training their model, generally with standard data augmentation (except for three participants), using various combinations of losses, most often including the Dice loss. The participants used various cross-validation schemes to optimize the generalization performance of their models. Half of the participants used an ensemble of multiple models.

4.3. Segmentation performance

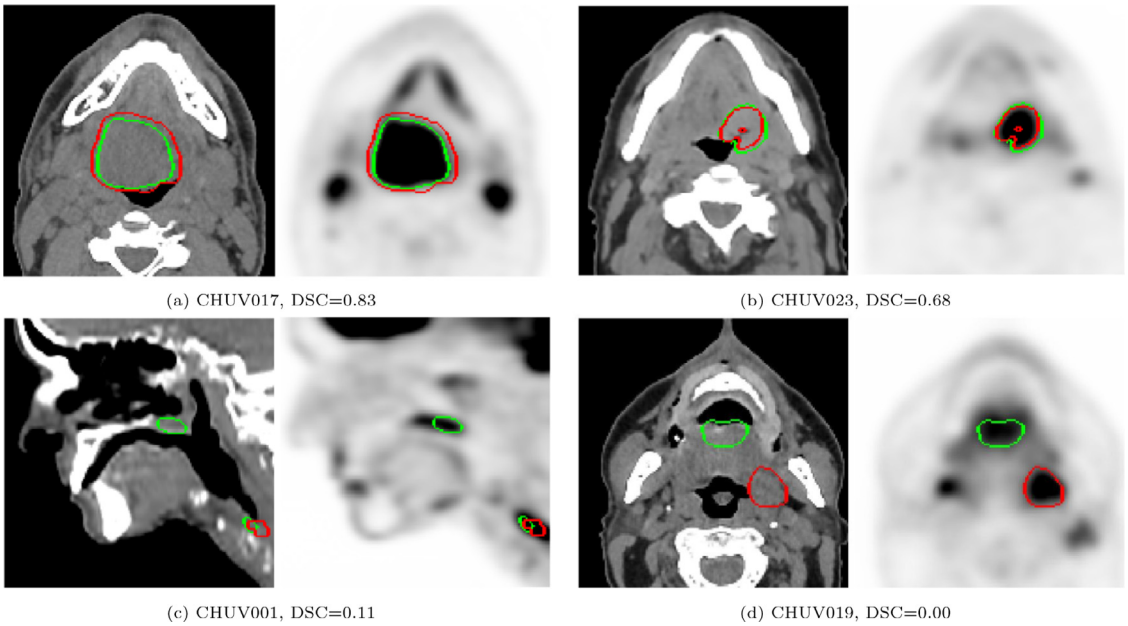
The results, including average DSC, precision, recall, and challenge rank are summarized in Table 3. We also report the average Surface Dice SCore at 1mm (SDSC) and the median Hausdorff Distance at 95% (HD95) as defined in (Nikolov et al., 2021). Our baseline method, developed in (Andrearczyk et al., 2020b) and provided to participants as an example on our GitHub repository, obtains an average DSC of 0.6588 and 0.6610 with the 2D and 3D implementations, respectively. Results on individual modalities are also reported for comparison.

The results from the participants (excluding post-challenge submissions) range from an average DSC of 0.5606 to 0.7591. (Iantsen et al., 2021b) (participant `andrei.iantsen`) obtained the best overall results with an average DSC of 0.7591, an average precision of 0.8332 and an average recall of 0.7400 (Fig. 2). This result (DSC) is not significantly higher than the second-best participant (Ma and Yang, 2021) (p-value of 0.3517 with a one-tailed Wilcoxon test). The statistical comparison of the score of each team is done in Fig. B.1 with the one-tailed Wilcoxon test and corrected for multiple hypotheses testing. Across all participants, the average precision ranges from 0.5850 to 0.8479. The recall ranges from 0.5022 to 0.8534, with the latter surprisingly obtained by the 3D PET/CT baseline (although with low precision, reflecting a trend to over-segment as compared to other algorithms). The median

[9] github.com/voreille/hecktor/tree/hecktor2020/src/evaluation, as of October 2021.

Table 3
Summary of the challenge results as of April 2021. The average DSC, precision, recall, SDSC and median HD95 are reported for the baseline algorithms and every team (the best result of each team). The unit of the HD95 is [mm]. The participant names are reported when no team name was provided. The ranking is only provided for teams that presented their method in a paper submission. The post-challenge results are denoted by an asterisk *. Bold values represent the best scores for each metric, excluding post-challenge results since we do not have any information about their method.

Team	DSC	HD95	Precision	Recall	SDSC	Rank
`paar`*	0.7624	3.27	0.8304	0.7490	0.6167	-
`andrei.iantsen` (Iantsen et al., 2021b)	**0.7591**	**3.27**	0.8333	0.7400	**0.6010**	1
`junma` (Ma and Yang, 2021)	0.7525	**3.27**	0.8384	0.7174	0.6003	2
`Fuller_MDA_Lab`*	0.7523	3.27	0.7838	0.7685	0.6168	-
`supratik_bose`*	0.7440	3.27	0.8350	0.7085	0.5822	-
`badger`*	0.7377	3.27	0.8143	0.7160	0.5800	-
`badger` (Xie and Peng, 2021)	0.7355	**3.27**	0.8326	0.7024	0.5735	3
`deepX` (Yuan, 2021)	0.7318	3.54	0.7851	0.7319	0.5528	4
`flash`*	0.7280	3.54	0.8020	0.7083	0.5650	-
`AIView_sjtu` (Chen et al., 2021)	0.7241	3.33	**0.8479**	0.6701	0.5598	5
`DCPT`	0.7049	4.10	0.7651	0.7047	0.5562	-
`xuefeng` (Ghimire et al., 2021)	0.6911	5.06	0.7525	0.6928	0.5011	6
`ucl_charp`	0.6765	5.42	0.7231	0.7257	0.5194	-
`QuritLab` (Yousefirizi and Rahmim, 2021)	0.6677	5.64	0.7289	0.7164	0.5086	7
`Unipa`	0.6674	4.10	0.7143	0.7039	0.4902	-
`Baseline 3D PET/CT`	0.6610	21.88	0.5909	**0.8534**	0.4502	-
`Baseline 2D PET/CT`	0.6588	26.81	0.6242	0.7629	0.4796	-
`HFHSegTeam` (Zhu et al., 2021)	0.6441	14.27	0.6938	0.6670	0.4922	8
`UESTC_501`	0.6382	5.16	0.6455	0.6874	0.4339	-
`Fuller_MDA_Lab` (Naser et al., 2021)	0.6373	5.06	0.7546	0.6283	0.4730	9
`Yone`*	0.6341	5.92	0.7690	0.6640	0.4513	-
`Baseline 3D PET`	0.6306	24.95	0.5768	0.8214	0.4399	-
`Baseline 2D PET`	0.6284	27.62	0.6470	0.6666	0.4231	-
`Maastro-Deep-L.` (Rao et al., 2021)	0.5874	29.56	0.6560	0.6142	0.4118	10
`Yone`	0.5737	21.46	0.6606	0.5590	0.4216	-
`SC_109`	0.5633	5.64	0.7652	0.5022	0.3542	-
`Roque`	0.5606	14.94	0.5850	0.6843	0.3601	-
`Baseline 2D CT`	0.3071	27.54	0.3477	0.3574	0.1847	-
`Baseline 3D CT`	0.2729	32.02	0.2154	0.5874	0.1218	-



(a) CHUV017, DSC=0.83 (b) CHUV023, DSC=0.68

(c) CHUV001, DSC=0.11 (d) CHUV019, DSC=0.00

Fig. 2. Examples of results of the winning algorithm (`andrei.iantsen` (Iantsen et al., 2021b)). The automatic segmentation results (green) and ground truth annotations (red) are displayed on 2D slices of PET (right) and CT (left) images. The reported DSC is computed on the entire image (see Eq. 1). (a), (b) Excellent segmentation results, detecting the GTVt of the primary oropharyngeal tumor localized at the base of the tongue and discarding the laterocervical lymph nodes despite high FDG uptake on PET. (c) Incorrect segmentation of the top volume at the level of the soft palate; (d) Incorrect segmentation of the smaller volume below the level of the hyoid bone. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

HD95 ranges from 3.27 to 32.02 [mm]. We chose to report the median since a value of $+\infty$ is attributed when the prediction is null. 3.27 [mm] is a highly observed value for HD95, which is probably due to the coarse axial resolution of the CT on the test set as we computed the performance in the original CT resolution (see C.1).

Note that two participants decided to withdraw their submissions due to very low scores. We allowed them to do so since their low scores were due to incorrect post-processing (*e.g.* setting incorrect pixel spacing or image origin), which was not representative of the performance of their algorithms. The distributions of DSCs across patients and across participants are reported in Figs. 3

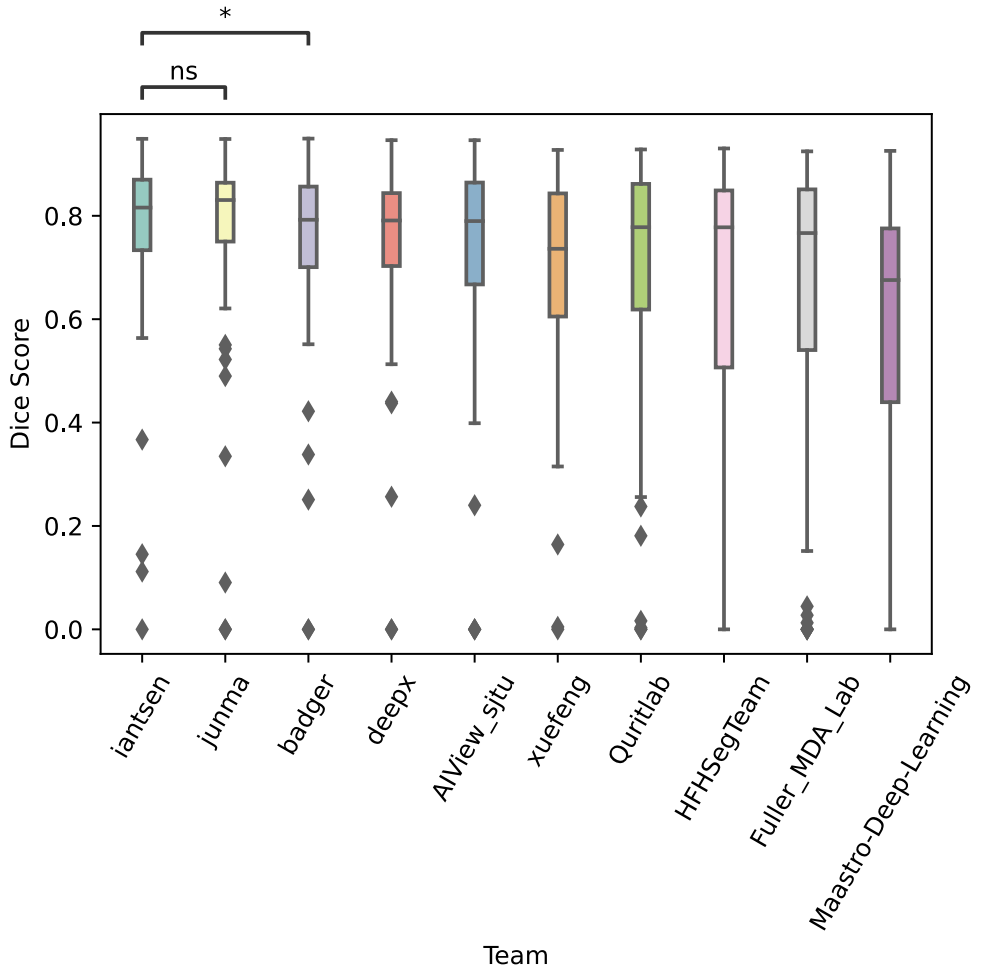


Fig. 3. Box plots of the distribution of the 53 test DSCs for each participant, ordered by decreasing rank.

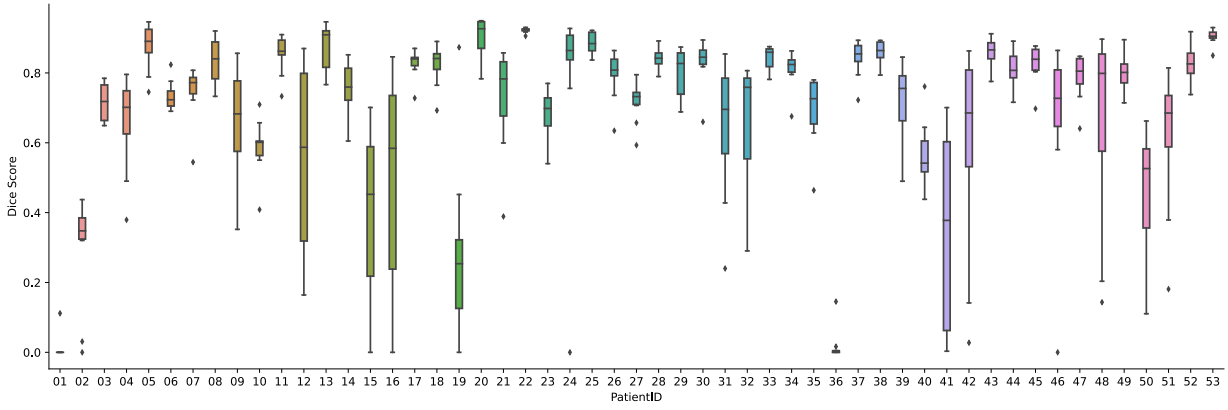


Fig. 4. Box plots of the distribution of DSCs across the 10 participants for each of the 53 patients in the test set.

and 4 respectively. Examples of segmentation results (TPs on top row, and FPs on bottom row) are shown in Fig. B.2.

4.4. Inter-observer agreement

We realized that it was crucial to also define the baseline for human observers performing the GTVt delineation task (*i.e.* segmentation), as well as their agreement. Three observers, *i.e.* two experts in radiation oncology and one nuclear physician, annotated the same 21 cases drawn randomly from the training and test sets and coming from all five centers. These 21 cases were chosen to represent approximately 10% of the dataset. It is worth noting that annotating the entire dataset four times was too costly. They were asked to delineate as close as possible the true tumoral volume as the aim is for radiomics studies. Together with the official challenge delineations, it amounts to four observers. All unique pairs of observers were considered, resulting in six pairs of comparisons. We computed the average DSC of all the pairs, *i.e.* all possible pairs of the four observers, which resulted in an average DSC of 0.6110. It is worth noting that for a faithful delineation of the tumor, a contrast-enhanced CT or an MRI image is required. Furthermore, there are no clinical guidelines for the task of segmenting GTVt

on PET/CT fusion. Moreover, the clinical information (*e.g.* physical examination) brings essential information to decide whether an abnormal structure is malignant. In this agreement, the observers were asked to perform this task with the PET/CT images only. Similar agreements were reported in the literature. (Gudi et al., 2017) reported the agreement of three observers with an average DSC of 0.57 using only the CT images for annotation and 0.69 using both PET and CT.

4.5. Ensemble of participants

In this section, we evaluate the possibility to ensemble the different participants' results into a "super-algorithm". Such analyses often revealed superior performances to all submitted runs (Menze et al., 2014), leveraging the diversity of the different methods (Hastie et al., 2009). We ensemble the (binary) predictions of all participants (with paper submissions, *i.e.* 10 participants) using the Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm (Warfield et al., 2004). This ensemble of predictions obtains an average DSC of 0.7574, a precision of 0.7301, and a recall of 0.8439. This result is better than the average performance of all participants (0.6931) and is slightly, but not significantly, outperformed by the best score of 0.7591 (p-value$=0.9230$). A simpler ensembling method is computed by taking the average of the 10 teams for each patient, and then, thresholding to 0.5 to obtain a binary prediction. This average prediction scores a DSC of 0.7426 which is not as good as the STAPLE ensembling (p-value$=0.044$). Note that several participants already reported results obtained as an ensemble of multiple independent network predictions. (see Table 2).

4.6. PET Thresholding

PET thresholding is *de facto* the most widely used method for lesion segmentation, at least in clinical routine, often via an initial manual delineation of the field of interest. As a comparison to the results obtained by the participants using deep learning automatic segmentation algorithms, we evaluate simple PET thresholding methods (automatic and semi-automatic). For the fully automatic threshold method, we simply threshold the PET image at a given percentage of the maximum SUV value within the bounding box.

For the semi-automatic threshold method, we mimic a manual indication of the GTVt followed by a threshold of the PET values. To this end, we threshold the PET image, compute the 26-connected components and retain the component that overlaps with the ground truth GTVt (or multiple components if more than one overlap with the ground truth GTVt). In Fig. 5, we report the results of both methods on the test set when varying the percentage of the maximum SUV used for thresholding. Finally, we also evaluate the same semi-automatic thresholding method with an additional threshold on the CT images (at -150 HU) to remove the air from the predictions. The best results, with an average DSC of 0.7409, are obtained with this semi-automatic PET/CT threshold at 30% of the maximum SUV value, which is aligned with previous findings, including in the context of the identification of predictive biomarkers (Castelli et al., 2017).

4.7. Tumor size and segmentation performance

In this section, we evaluate how the algorithms perform for different tumor sizes. To this end, we explore the correlation of tumor size with the performance of the algorithms. The tumor size is calculated as the voxel count inside the ground truth GTVt multiplied by the voxel volume. The Spearman correlation across all ten participants and all tumors is 0.4301 (p-value< 0.001). In Fig. 6, we illustrate this correlation with a scatter plot of the DSC as a function of tumor size. Fig. 7 relates the performance for each of the 10 algorithms for four tumor size groups. This figure was generated by grouping the 53 test cases in 4 bins (*i.e.* intervals) of 13, 13, 13, and 14 cases, respectively. The average DSC was then computed for each team in each bin.

4.8. Analysis of false positives

In this section, we want to evaluate, for a given algorithm, whether FPs are generally occurring in the surroundings of the ground truth GTVt, or biased towards other regions with high FDG

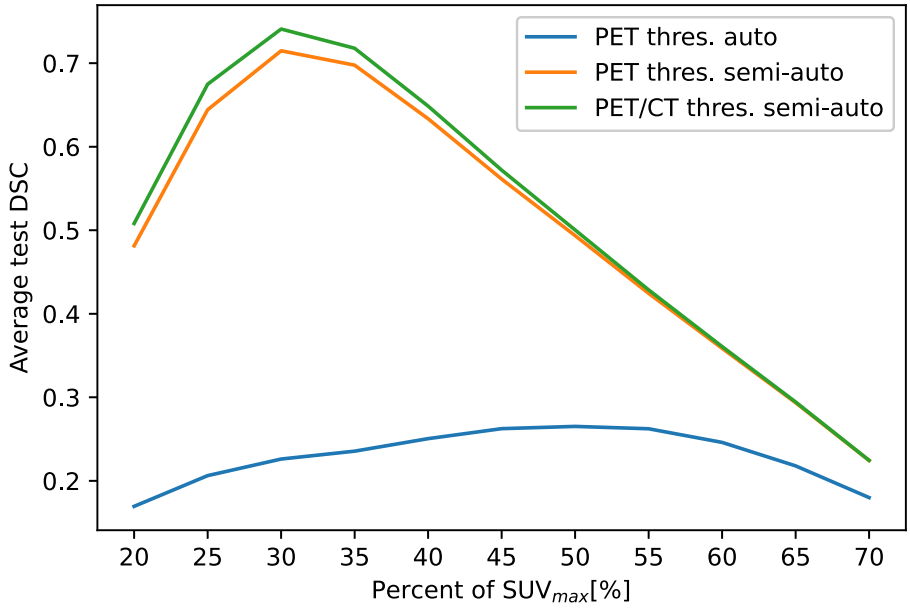


Fig. 5. Segmentation performance of PET thresholding-based method at different percentages of maximum SUV. Three results are reported: the automatic PET threshold, the semi-automatic PET threshold (indicating the location of the ground truth GTVt), and the semi-automatic PET and CT (for removing the air) threshold.

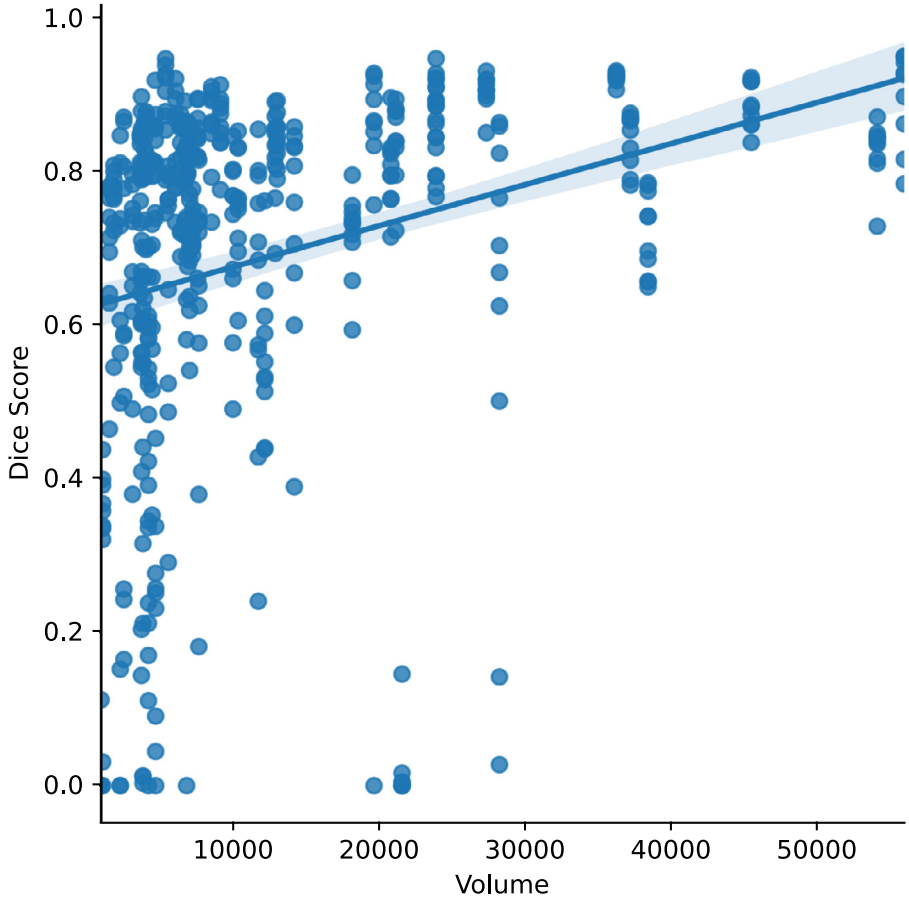


Fig. 6. Scatter plot of DSC vs. tumor volume (voxel count in the VOI) for 10 participants. The corresponding Spearman correlation is 0.43.

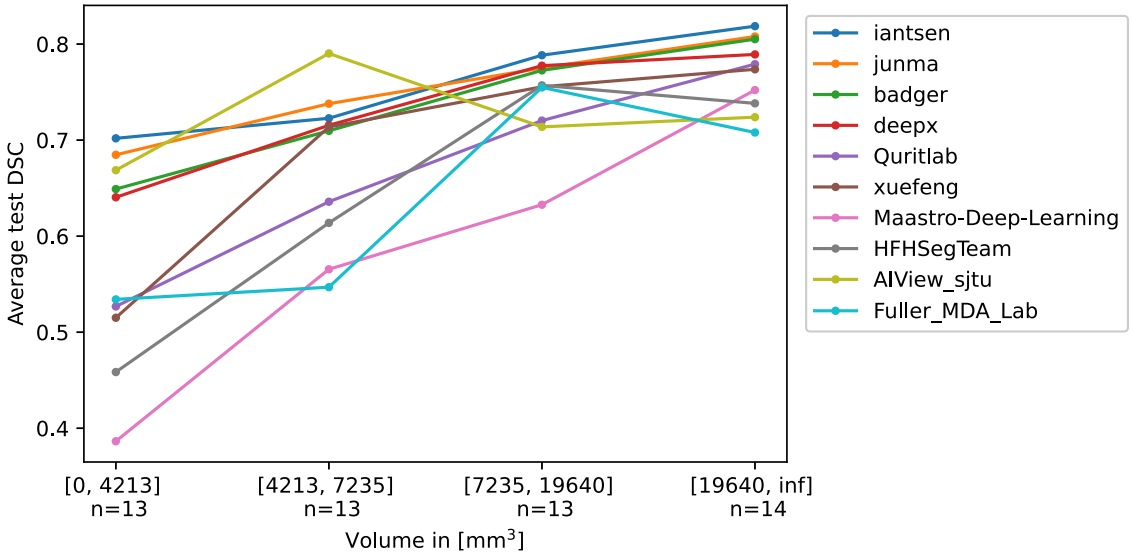
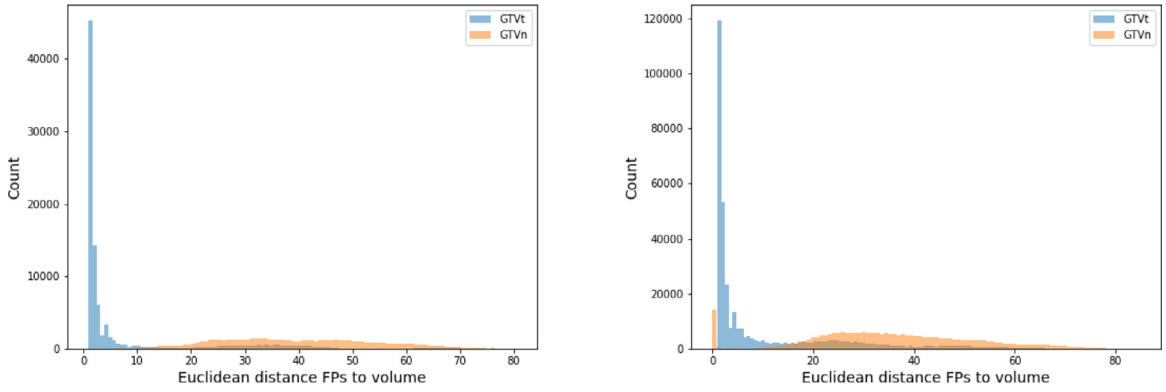


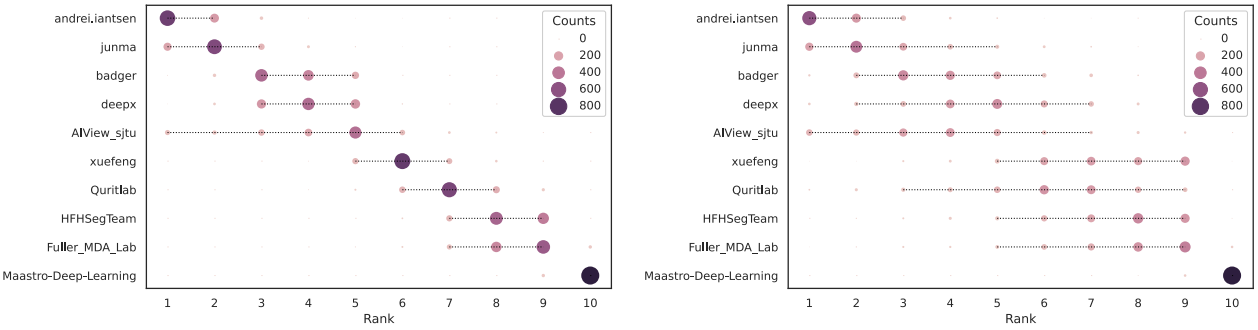
Fig. 7. Average DSC of each team's algorithm in function of the volume of the tumors. This figure was generated by distributing the 53 test volumes in 4 bins of n =13, 13, 13, and 14 each and then computing the average DSC for each bin.

uptakes such as the lymph nodes or other zones with inflammation. To this end, we compute the shortest Euclidean distance of each FP voxel to the ground truth GTVt. We then aggregate these distances for all test cases and report these values into a histogram in Fig. 8. Similarly, we compute the distance of each FP voxel to the ground truth GTVn (lymph nodes) and report the histogram on the same figure. We compute this analysis for the best participant (`andrei.iantsen`), as well as for the baseline (3D PET/CT U-Net) since it was the approach with the largest recall but low precision. Note that we only compute the histogram of the FP voxels to avoid squashing the counts of the non-zero bins due to the large number of TPs with a distance to the GTVt of zero (first bin).



(a) `andrei.iantsen` (b) baseline 3D PET/CT

Fig. 8. Histogram of the Euclidean distance of the FP voxels to the closest ground truth GTVt voxel and GTVn voxel. We evaluate here the prediction of the first ranked participant (`andrei.iantsen`) (a) and our baseline 3D PET/CT (b). For comparison, the False Discovery Rate (FDR), i.e. FP/(FP+TP) is 0.15, with 544,343 TPs in (a) and FDR = 0.37 with 621,413 TPs in (b).



(a) Rank based on average DSC (b) (Alternative) rank based on average rank DSC

Fig. 9. Ranking robustness against changes in test data. The robustness is assessed by ranking 1000 bootstraps of the test set. The size of the circles is proportional to the number of times a team obtained the corresponding rank for each bootstrap. The dashed lines represent the confidence intervals at 95% computed from the bootstrap analysis. The current ranking, *i.e.* the one used in this challenge, is obtained by averaging the DSCs across all test cases. The alternative ranking is computed by averaging the rankings of each team across the test cases.

4.9. Ranking robustness

Ranking robustness against changes in the test set is assessed by evaluating the variation of the ranking on 1000 bootstrap repetitions of the test set. We also compared the current ranking against an alternate ranking defined as follows. This alternative ranking was computed based on the average ranking across all cases. If multiple teams obtain the same rank for one case, the average rank is attributed to these teams. For instance, if three participants score 0 on a given case, the average rank of $\frac{8+9+10}{10} = 9$ is attributed to all of them for this case.

Fig. 9 depicts the results of the bootstrap analysis for the two rankings. We also computed the Kendall rank correlation coefficient between the ranking of each bootstrap and the ranking on the whole test set. We obtained 0.8772 (0.7333 - 1.0000) and 0.7335 (0.4658 - 0.9111) for the current ranking and alternate ranking, respectively. The numbers in parenthesis are the confidence intervals at 95% computed with the bootstrap eanalysis. The methodology used in this section to report ranking robustness is inspired by the challengeR toolkit (Wiesenfarth et al., 2021).

5. Discussion

This section interprets and discusses the results reported in Section 4. We first discuss and report the overall challenge participation and main lessons learned. Second, the segmentation performance achieved by all participating methods is interpreted. Finally, we report the current limitations and sources of errors of this challenge.

5.1. Participation and main lessons learned

This challenge allowed us to compare state-of-the-art algorithms developed by 18 teams across the world on the task of primary H&N tumor segmentation in PET/CT images. Excellent results were obtained with the first ranked team reaching 0.7591 average DSC, 0.8332 precision, and 0.7400 recall. In Table 2, we attempted to group the results based on important elements of the algorithms. In particular, we identified several elements important for addressing the task. All participants used U-Net based architectures, mostly 3D. Preprocessing, normalization, data augmentation, and ensembling seem to play an important role in the final results. Most of these trends (see also algorithms description in Section 4.2) and results can be found in other medical imaging segmentation challenges (Menze et al., 2014; Ma, 2021). An interesting comparison of several challenges (including HECKTOR 2020) and algorithms focusing on automatic segmentation in medical images can be found in (Ma, 2021).

We note, however, that it is a difficult task to characterize algorithms with only a few descriptions and to assign good performance to specific parts. The methods are highly complex with high degrees of freedom and many hyper-parameters that can all have

a strong influence on segmentation performance. Simple modifications such as the number of training iterations or the learning rate can have a large impact on the results and cannot be exhaustively listed and compared. For this analysis, we asked the participants to specifically report a set of characteristics of their algorithms to be able to compare them in Table 2. More information will be asked in the future editions of HECKTOR to enhance comparison.

The ranking used in this edition was based on the average DSC. The results of Section 4.9 show that this approach is more robust to changes in the test set. These findings are corroborated by Maier-Hein et al. (2018) where they showed that ranking based on averaged metrics are more consistent for changes in test data.

5.2. Overall segmentation performance

As shown in Fig. 4, some cases were incorrectly segmented by most or all participants, *e.g.* CHUV01 and CHUV36. On the contrary, some cases were correctly segmented by most participants (*e.g.* CHUV22 and CHUV53), and others showed a large variability across participants' algorithms (*e.g.* CHUV16 and CHUV41). These differences, as confirmed by further evaluations in Sections 4.8, 4.7, originate from the tumor size, the SUVs within the GTVt, and the presence of lymph nodes or other regions with high SUVs. Some examples are illustrated in Fig. B.2.

The participants' algorithms obtained better results than the inter-observer agreement. This comparison, however, should be put into perspective. First, the cases used in the agreement were different from the test set. Second, one annotator, the one who annotated the entire dataset for the challenge, had extra information since he corrected the radiotherapy annotations whereas the others were asked to draw the segmentation from scratch without any further information than the raw PET/CT data. Finally, some annotators delineated closer to radiotherapy requirements, *i.e.* with large annotations, resulting in higher disagreement. To alleviate this issue, we are currently developing clear guidelines for the next iteration of the challenge.

The results can also be compared with a simple PET thresholding method (see Section 4.6), often used in radiomics studies (Erdi et al., 1997; Castelli et al., 2017). The latter obtained an average DSC of 0.7409 when used in a semi-automatic manner. This result is significantly lower than the performance of the best participants (0.7591, p-value of 0.0237) and must be considered with precaution since the segmentation was highly guided toward the true tumor location and the threshold was optimized on the test set. With a fully automatic threshold of the PET image in the oropharynx region, we only obtain 0.2652 due to various regions, including lymph nodes, with high SUVs. The best semi-automatic threshold method was obtained with a threshold around 30% of the maximum SUV, as frequently used to measure the metabolic response characteristic of the tumor, *e.g.* 36–44% for best approximation of tumor volume (Erdi et al., 1997), 40 to 68% of SUV max for best radiomics results in DFS prediction (Castelli et al., 2017; Creff et al., 2020). Overall, this suggests that the segmentation algorithms can leverage the wealth of both PET and CT images (*i.e.* metabolic and anatomical/structural tumor properties) to provide more advanced segmentation rules when compared to simple PET thresholding. This is also corroborated by the consistent superiority of algorithms using both PET and CT imaging modalities when compared to using PET only.

The ensemble of participants' methods (see Section 4.5) reached a good consensus with an average DSC of 0.7574 and a rather high recall (0.8438) and low precision (0.7301) as compared to other results in the same range. While this is not better than the first rank result, it would likely achieve an excellent generalization to other data.

5.3. Detailed performance analysis

The analysis of tumor sizes in Section 4.7 (Figs. 6 and 7) showed that they are correlated with the segmentation performance. These results seem to show that the small tumor sizes are more difficult to segment than the large ones. More precisely, smaller tumors are less consistently well segmented, resulting in a large variation of performance. This is not surprising since small lesions suffer from a higher partial volume effect which increases the relative difficulty to define the boundary of the tumor (Foster et al., 2014). Moreover, the volumetric (or 3D) DSC is largely dependent on the volume sizes. A contour deviation of $\pm$1mm around the true tumor boundary, for instance, will affect DSC values more for small tumors than the large ones, resulting in a negligible chance for the latter.

In Fig. 8 (Section 4.8), we analyzed the spatial arrangement of FPs segmented voxels. We conducted this experiment for the first ranked results and our baseline. In both cases, the majority of FP voxels are located in the surrounding of the GTVt, as shown in Fig. 8. As illustrated in the same figure, the FPs of the best results are not located near the lymph nodes, whereas a lot of FPs of the baseline are located in the lymph nodes and their surroundings. This suggests that, unlike the baseline, the best algorithm relies on true tumoral patterns and not only on FDG uptake.

5.4. Limitations and sources of errors

The main limitation of the current challenge is the lack of more precise GTVt ground truth. The annotations were made on the PET/CT fusion without using other modalities such as contrast-enhanced CT or MR which allow delineating the tumor more faithfully. This limitation is illustrated by the results of the inter-observer agreement mentioned in Section 4.4, where the average DSC of 0.6110 highlighted the difficulty of the task. A source of error, therefore, originates from the degree of subjectivity and the lack of guidelines in the annotation and correction of the expert.

Another limitation of this challenge is the lack of test data with exact ground truth. To obtain such data, phantom and simulation can be used. This enables the evaluation of performances of models on data where the exact ground truth is known. (Hatt et al., 2017) claim that for a good benchmark in PET segmentation, one must include simulated and phantom test images in addition to clinical test data.

In this challenge, we provided the participants with a bounding box to decrease the difficulty of the task. This can be seen as a limitation since the resulting methods are not fully automatic, but these bounding boxes cover a large portion of the original image and are easy to detect automatically (Andrearczyk et al., 2020a).

6. Conclusions

This paper presents the HECKTOR 2020 challenge on the segmentation of the primary tumor of oropharyngeal H&N cancer in FDG PET/CT. Detailed information was reported on the dataset, participation, and segmentation performance. Good participation with 18 teams and 10 participants' publications allowed us to compare state-of-the-art segmentation methods on this challenging task. The results are very satisfactory with the winning team achieving an average DSC of 0.7591, which is superior to the inter-observer agreement (average DSC 0.6110). These results were obtained with a strict testing scheme as the test cases were all from an unseen center. It is reasonable to expect better results if the proposed methods are fine-tuned on few examples from this center. All participants used U-Net based deep learning models, most of them

with a 3D architecture and standard pre-processing techniques. We could identify several key elements that seem to have led to good results, including normalization, data augmentation, and ensembling of multiple models.

Preliminary experiments show that fully automatic radiomics methods are on pair or surpass radiomics models based on feature extraction from manual annotations (Fontaine et al., 2021; Andrearczyk et al., 2021a). These preliminary results are very encouraging and demonstrate that we are one step closer to analyzing very large-scale cohorts for radiomics validation.

While focusing on H&N cancer in HECKTOR, we believe that many of the methods developed and lessons learned will generalize to the automatic segmentation of other types of cancer imaged in PET/CT images (*e.g.* lung, melanoma).

In future editions, we aim to increase the size of the dataset and propose other clinically relevant tasks such as the segmentation of lymph nodes and the prediction of patient outcome (*e.g.* disease-free survival).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Valentin Oreiller: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Writing – original draft, Writing – review & editing. **Vincent Andrearczyk:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Writing – original draft, Writing – review & editing. **Mario Jreige:** Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Writing – review & editing. **Sarah Boughdad:** Conceptualization, Data curation, Validation, Writing – review & editing. **Hesham Elhalawani:** Conceptualization, Data curation, Validation, Writing – review & editing. **Joel Castelli:** Conceptualization, Data curation, Validation, Writing – review & editing. **Martin Vallières:** Conceptualization, Data curation, Methodology, Writing – review & editing. **Simeng Zhu:** Investigation, Writing – review & editing. **Juanying Xie:** Investigation, Writing – review & editing. **Ying Peng:** Investigation, Writing – review & editing. **Andrei Iantsen:** Investigation, Writing – review & editing. **Mathieu Hatt:** Investigation, Writing – review & editing. **Yading Yuan:** Investigation, Writing – review & editing. **Jun Ma:** Investigation, Writing – review & editing. **Xiaoping Yang:** Investigation, Writing – review & editing. **Chinmay Rao:** Investigation, Writing – review & editing. **Suraj Pai:** Investigation, Writing – review & editing. **Kanchan Ghimire:** Investigation, Writing – review & editing. **Xue Feng:** Investigation, Writing – review & editing. **Mohamed A. Naser:** Investigation, Writing – review & editing. **Clifton D. Fuller:** Investigation, Writing – review & editing. **Fereshteh Yousefirizi:** Investigation, Writing – review & editing. **Arman Rahmim:** Investigation, Writing – review & editing. **Huai Chen:** Investigation, Writing – review & editing. **Lisheng Wang:** Investigation, Writing – review & editing. **John O. Prior:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Adrien Depeursinge:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Acknowledgments

We thank all the teams for their valuable work. This challenge and the winner prize were kindly sponsored by Siemens Healthineers Switzerland. This work was also partially supported by the Swiss National Science Foundation (SNSF, grant 205320_179069) and the Swiss Personalized Health Network (SPHN, via the IMAGINE and QA4IQI projects).

Appendix A. Participants' Algorithms Summary

In (Iantsen et al., 2021b), Iantsen et al. proposed a model based on a U-Net architecture with residual layers and supplemented with 'Squeeze and Excitation' (SE) normalization, previously developed by the same authors for brain tumor segmentation. An unweighted sum of soft Dice loss and Focal Loss was used for training. The test results were obtained as an ensemble of eight models trained and validated on different splits of the training set. No data augmentation was performed.

In (Ma and Yang, 2021), Ma and Yang used a combination of U-Nets and hybrid active contours. First, 3D U-Nets are trained to segment the tumor (with a cross-validation on the training set). Then, the segmentation uncertainty is estimated by model ensembles on the test set to select the cases with high uncertainties. Finally, the authors used a hybrid active contour model to refine the high uncertainty cases. The U-Nets were trained with an unweighted combination of Dice loss and top-K loss. No data augmentation was used.

In (Zhu et al., 2021), Zhu et al. used a two steps approach. First, a classification network (based on ResNet) selects the axial slices which may contain the tumor. These slices are then segmented using a 2D U-Net to generate the binary output masks. Data augmentation was applied by shifting the crop around the provided bounding boxes and the U-Net was trained with a soft Dice loss. The preprocessing includes clipping the CT and the PET, standardizing the HU within the cropped volume and scaling the range of the PET to correspond to the CT range by dividing it by a factor of 10.

In (Yuan, 2021), Yuan proposed to integrate information across different scales by using a dynamic Scale Attention Network (SA-Net), based on a U-Net architecture. Their network incorporates low-level details with high-level semantics from feature maps at different scales. The network was trained with standard data augmentation and with a Jaccard distance loss, previously developed by the authors. The results on the test set were obtained as an ensemble of ten models.

In (Chen et al., 2021), Chen et al. proposed a three-step framework with iterative refinement of the results. In this approach, multiple 3D U-Nets are trained one-by-one using a Dice loss without data augmentation. The predictions and features of previous models are captured as additional information for the next one to further refine the segmentation.

In (Ghimire et al., 2021), Ghimire et al. developed a patch-based approach to tackle the memory issue associated with 3D images and networks. They used an ensemble of conventional convolutions (with small receptive fields capturing fine details) and dilated convolutions (with a larger receptive field of capturing global information). They trained their model with a weighted cross-entropy and dice loss and random left-right flips of the patches were applied for data augmentation. Finally, an ensemble of the best two models selected during cross-validation was used for predicting the segmentation of the test data.

In (Yousefirizi and Rahmim, 2021), Yousefirizi and Rahmim proposed a deep 3D model based on SegAN, a generative adversarial network (GAN) for medical image segmentation. An improved polyphase V-net (to help preserve boundary details) is used for the generator and the discriminator network has a similar structure to the encoder part of the former. The networks were trained using a combination of Mumford-Shah (MS) and multi-scale Mean Absolute Error (MAE) losses, without data augmentation.

In (Xie and Peng, 2021), Xie and Peng proposed a 3D scSE nnU-Net model, improving upon the 3D nnU-Net by integrating the spatial and channel 'Squeeze and Excitation' (scSE) blocks. They trained the model with a weighted combination of Dice and cross-entropy losses, together with standard data augmentation techniques (rotation, scaling etc.). To preprocess the CT images an automated level-window-like clipping of intensity values is performed based on the 0.5 and 99.5th percentile of these values. The intensity values of the PET are standardized by subtracting the mean and then, by dividing by the standard deviation of the image.

In (Naser et al., 2021), Naser et al. used a variant of 2D and 3D U-Net (we report the best result, with the 3D model). The models were trained with a combination of Dice and cross-entropy losses with standard data augmentation.

In (Rao et al., 2021), Rao et al. proposed an ensemble of two methods, namely a 3D U-Net and another 2D U-Net variant with 3D context. A top-k loss was used to train the models without data augmentation.

Appendix B. Additional plots

This appendix presents additional plots. In Fig. B.1 the pairwise statistical comparison of the 10 teams is illustrated by a significance matrix computed with a corrected one-sided Wilcoxon signed-rank test at 5% significance. In Fig. B.2, examples of predictions obtained by the second-ranked team (`junma` (Ma, Yang, 2021)) are drawn on the same cases as Fig. 2 to illustrate the variability among the two best teams. Figs. B.3, B.4 and B.5 show, for each participant, the distributions across the 53 test cases of the precision, recall, and SDSC, respectively.

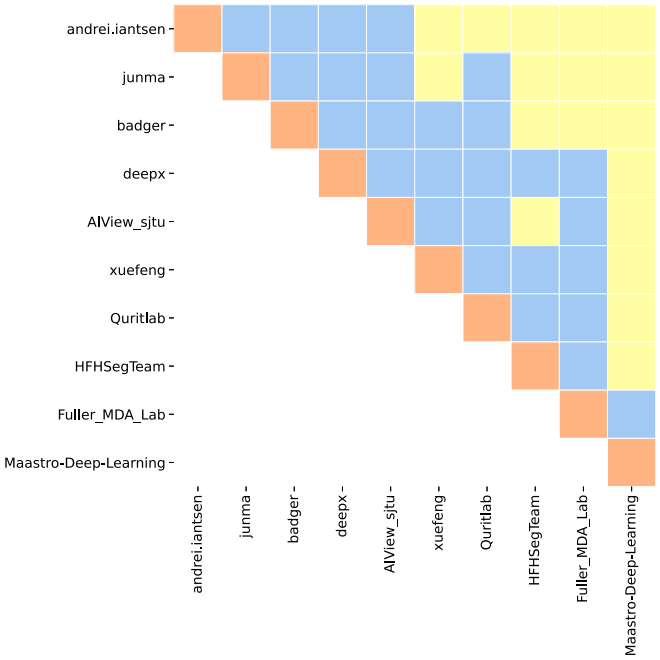
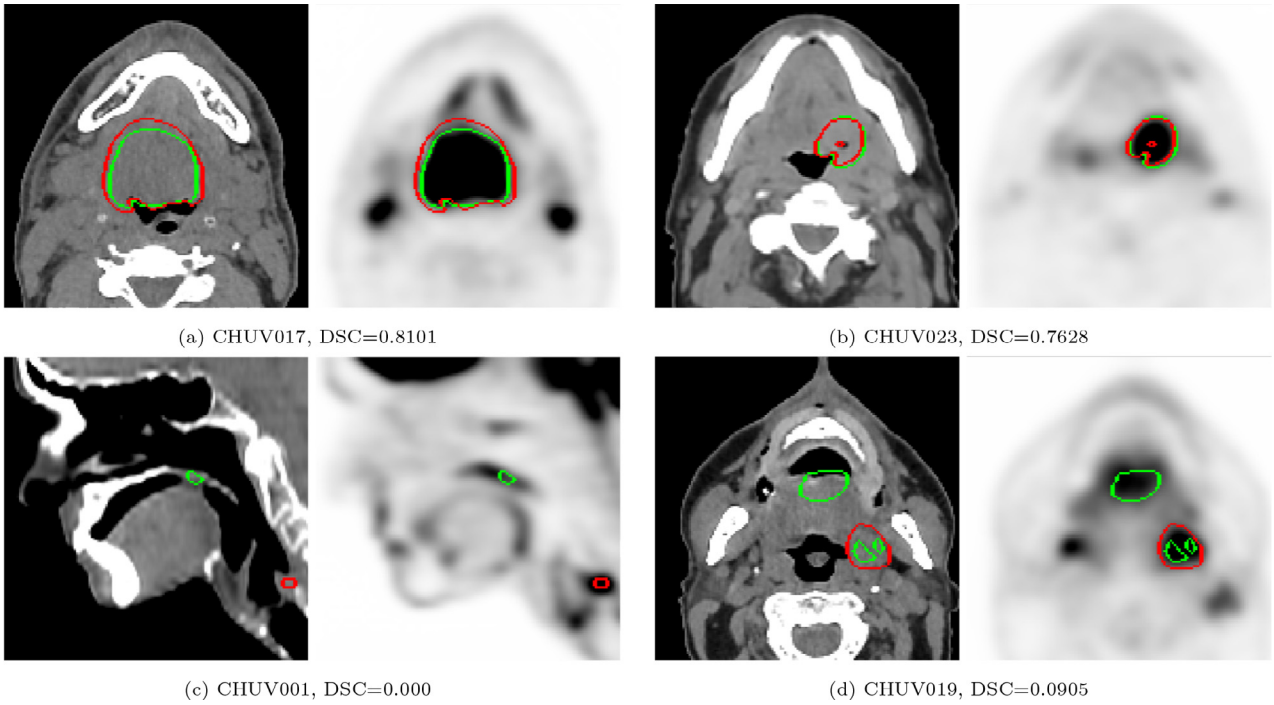


Fig. B.1. The significance matrix represents significant tests for the one-sided Wilcoxon signed-rank test at a 5% significance level, adjusted for multiple comparisons with the Holm-Bonferroni method for 45 hypotheses. For each pair, the alternative hypothesis is that the best team has a greater score. For instance, for the `andrei.iantsen-junma` pair the alternative is that `andrei.iantsen` has a better DSC than `junma`. The yellow color indicates that the team on the line of the matrix has significantly better DSC than the team on the column. Blue color means no significant difference. Orange color is used as a visual guide to show pairs of identical teams. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) CHUV017, DSC=0.8101 (b) CHUV023, DSC=0.7628

(c) CHUV001, DSC=0.000 (d) CHUV019, DSC=0.0905

Fig. B.2. Examples of results of the second algorithm (`junma` (Ma and Yang, 2021)). The automatic segmentation results (green) and ground truth annotations (red) are displayed on 2D slices of PET (right) and CT (left) images. The reported DSC is computed on the entire image (see Eq. 1). (a), (b) Excellent segmentation results, detecting the GTVt of the primary oropharyngeal tumor localized at the base of the tongue and discarding the laterocervical lymph nodes despite high FDG uptake on PET. (c) Incorrect segmentation of the top volume at the level of the soft palate; (d) Incorrect segmentation of the smaller volume below the level of the hyoid bone. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

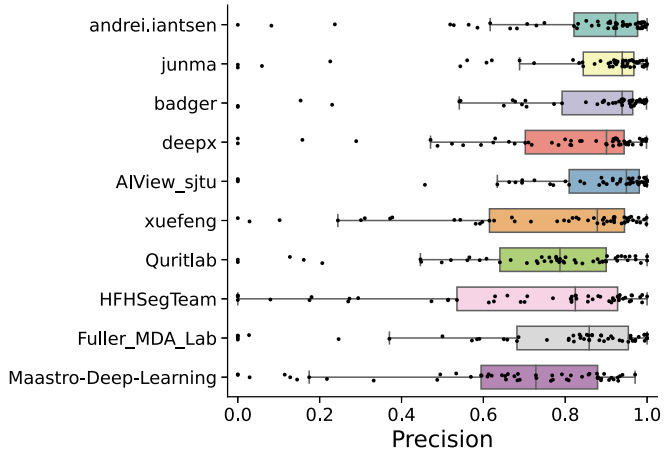


Fig. B.3. Box plots of the distribution of the precision on the 53 test cases for each participant, ordered by decreasing rank.

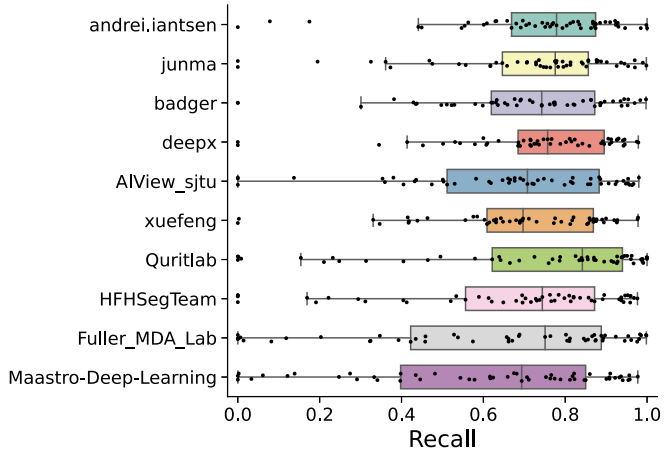


Fig. B.4. Box plots of the distribution of the recall on the 53 test cases for each participant, ordered by decreasing rank.

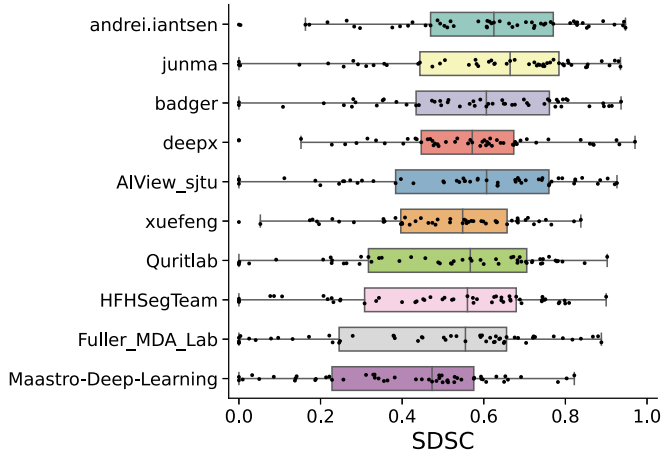


Fig. B.5. Box plots of the distribution of the 53 test SDSCs for each participant, ordered by decreasing rank.

Appendix C. Centers statistics

In Table C.1, we report the differences between the five centers in terms of image properties such as devices, pixel spacing and slice spacing. We also disclose the distribution of GTVt volumes in Fig. C.1 and Table C.2

Table C.2
Average GTVt volume for the five center used in this challenge. The numbers in parenthesis represent the 5th and 95th respectively.

Center	GTVt volume
HGJ	14.913 (2.263 - 38.879)
CHUS	14.209 (1.837 - 42.967)
HMR	23.622 (2.412 - 88.785)
CHUM	9.866 (1.358 - 24.884)
CHUV	13.317 (1.725 - 41.212)

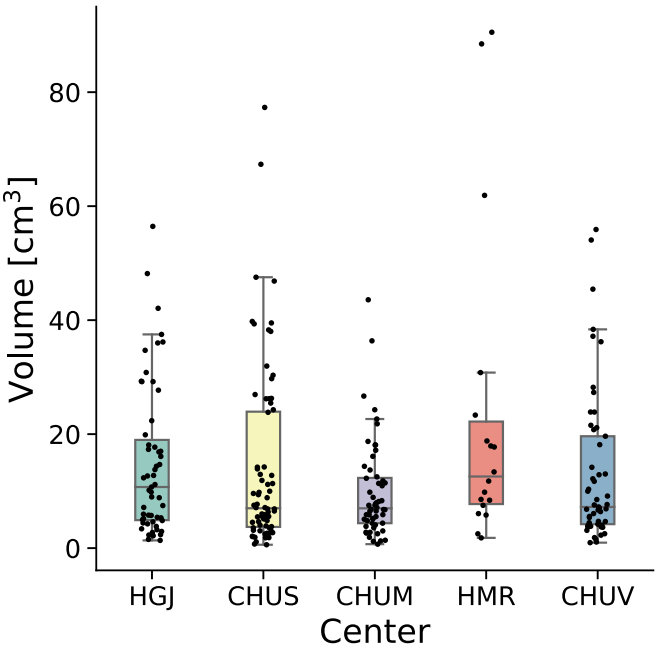


Fig. C.1. Box plots of the distribution of the GTVt volumes per center.

Table C.1
Statistics of the different centers. GTVt volumes are computed after isoresampling at $1 \times 1 \times 1$ mm^3. The GTVt volumes are reported in cm^3 as average plus the 5th and 95th percentile in parenthesis. All devices are hybrid PET/CT.

Center	Pixel spacing CT	Slice spacing CT	Pixel spacing PT	Slice spacing PT	Device
HGJ	0.98 (0.98 - 0.98)	3.27 (3.27 - 3.27)	3.52 (3.52 - 4.69)	3.27 (3.27 - 3.27)	Discovery ST, GE Healthcare
CHUS	1.17(0.68- 1.17)	3.00 (2.00 - 5.00)	4.00 (4.00 - 4.00)	4.00 (4.00 - 4.00)	GeminiGXL 16, Philips
HMR	0.98 (0.98 1.37)	3.27 (3.27 - 3.27)	3.52 (3.52 - 5.47)	3.27 (3.27 - 3.27)	Discovery STE, GE Healthcare
CHUM	0.98 (0.98 - 1.37)	1.50 (1.50 - 3.27)	4.00 (3.52 - 5.47)	4.00 (3.27 - 4.06)	Discovery STE, GE Healthcare
CHUV	1.37 (0.98 - 1.37)	3.27 (1.00 - 4.25)	2.73 (2.73 - 3.91)	3.27 (3.27 - 4.25)	Discovery D690 TOF, GE Healthcare

References

Andrearczyk, V., Fontaine, P., Oreiller, V., Castelli, J., Jreige, M., O. Prior, J., Depeursinge, A., 2021. Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer. PRedictive Intelligence in MEdicine (PRIME at MICCAI).

Andrearczyk, V., Oreiller, V., Depeursinge, A., 2020. Oropharynx detection in PET-CT for tumor segmentation. Irish Machine Vision and Image Processing.

Andrearczyk, V., Oreiller, V., Vallières, M., Castelli, J., Elhalawani, H., Jreige, M., Boughdad, S., Prior, J.O., Depeursinge, A., 2020. Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In: International Conference on Medical Imaging with Deep Learning (MIDL).

Andrearczyk, V., Oreiller, V., Vallières, M., Jreige, M., Prior, J.O., Depeursinge, A., 2021. Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. Lecture Notes in Computer Science (LNCS) Challenges.

Aristophanous, M., Penney, B.C., Martel, M.K., Pelizzari, C.A., 2007. A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography. Med. Phys. 34 (11), 4223–4235.

Blanc-Durand, P., Van Der Gucht, A., Schaefer, N., Itti, E., Prior, J.O., 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, M., Riesterer, O., Stark, L.S., Studer, G., Unkelbach, J., Guckenberger, M., Tanadini-Lang, S., 2017. Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. Acta Oncol. 56 (11), 1531–1536.

Bonner, J.A., Harari, P.M., Giralt, J., Cohen, R.B., Jones, C.U., Sur, R.K., Raben, D., Baselga, J., Spencer, S.A., Zhu, J., 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. Lancet Oncol. 11 (1), 21–28.

Castelli, J., Depeursinge, A., De Bari, B., Devillers, A., De Crevoisier, R., Bourhis, J., Prior, J.O., 2017. Metabolic tumor volume and total lesion glycolysis in oropharyngeal cancer treated with definitive radiotherapy: which threshold is the best predictor of local control? Clin. Nucl. Med. 42 (6), e281–e285.

Chajon, E., Lafond, C., Louvel, G., Castelli, J., Williaume, D., Henry, O., Jégoux, F., Vauléon, E., Manens, J.-P., Le Prisé, E., 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. Radiat. Oncol. 8 (1), 132.

Chen, H., Chen, H., Wang, L., 2021. Iteratively refine the segmentation of head and neck tumor in FDG-PET and CT images. Lecture Notes in Computer Science (LNCS) Challenges.

Creff, G., Devillers, A., Depeursinge, A., Palard-Novello, X., Acosta, O., Jegoux, F., Castelli, J., 2020. Evaluation of the prognostic value of FDG PET/CT parameters for patients with surgically treated head and neck cancer: asystematic review. JAMA Otolaryngol. Head Neck Surg. 146 (5), 471–479.

del Toro, O.A.J., Goksel, O., Menze, B., Müller, H., Langs, G., Weber, M.-A., Eggel, I., Gruenberg, K., Holzer, M., 2014. VISCERAL–VISual Concept Extraction Challenge in RAdioLogy: ISBI 2014 Challenge Organization.

Erdi, Y.E., Mawlawi, O., Larson, S.M., Imbriaco, M., Yeung, H., Finn, R., Humm, J.L., 1997. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. Cancer 80 (S12), 2505–2509.

Fontaine, P., Andrearczyk, V., Oreiller, V., Castelli, J., Jreige, M., O. Prior, J., Depeursinge, A., 2021. Fully automatic head and neck cancer prognosis prediction in PET/CT. Multimodal Learning and Fusion Across Scales for Clinical Decision Support (ML-CDS at MICCAI).

Foster, B., Bagci, U., Mansoor, A., Xu, Z., Mollura, D.J., 2014. A review on segmentation of positron emission tomography images. Comput. Biol. Med. 50, 76–96.

Fu, X., Bi, L., Kumar, A., Fulham, M., Kim, J., 2021. Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. IEEE J. Biomed. Health Inform..

Ghimire, K., Chen, Q., Feng, X., 2021. Patch-based 3D UNet for head and neck tumor segmentation with an ensemble of conventional and dilated convolutions. Lecture Notes in Computer Science (LNCS) Challenges.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al., 2018. Niftynet: a deep-learning platform for medical imaging. Comput. Methods Programs Biomed. 158, 113–122.

Gillies, R.J., Kinahan, P.E., Hricak, H., 2016. Radiomics: images are more than pictures, they are data. Radiology 278 (2), 563–577.

Groendahl, A.R., Knudtsen, I.S., Huynh, B.N., Mulstad, M., Moe, Y.M., Knuth, F., Tomic, O., Indahl, U.G., Torheim, T., Dale, E., et al., 2021. A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. Phys. Med. Biol. 66 (6), 065012.

Gudi, S., Ghosh-Laskar, S., Agarwal, J.P., Chaudhari, S., Rangarajan, V., Paul, S.N., Upreti, R., Murthy, V., Budrukkar, A., Gupta, T., 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. J. Med. Imaging Radiat. Sci. 48 (2), 184–192.

Guo, Z., Guo, N., Gong, K., Li, Q., et al., 2019. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys. Med. Biol. 64 (20), 205015.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Series in Statistics. Springer New York, New York, NY.

Hatt, M., Laurent, B., Ouahabi, A., Fayad, H., Tan, S., Li, L., Lu, W., Jaouen, V., Tauber, C., Czakon, J., Drapejkowski, F., Dyrka, W., Camarasu-Pop, S., Cervenansky, F., Girard, P., Glatard, T., Kain, M., Yao, Y., Barillot, C., Kirov, A., Visvikis, D., 2018. The first MICCAI challenge on PET tumor segmentation. Med. Image Anal. 44, 177–195.

Hatt, M., Le Rest, C.C., Turzo, A., Roux, C., Visvikis, D., 2009. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans. Med. Imaging 28 (6), 881–893.

Hatt, M., Lee, J.A., Schmidtlein, C.R., Naqa, I.E., Caldwell, C., De Bernardi, E., Lu, W., Das, S., Geets, X., Gregoire, V., et al., 2017. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM task group no. 211. Med. Phys. 44 (6), e1–e42.

Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al., 2021. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. Med. Image Anal. 67, 101821.

Huang, B., Chen, Z., Wu, P.-M., Ye, Y., Feng, S.-T., Wong, C.-Y.O., Zheng, L., Liu, Y., Wang, T., Li, Q., et al., 2018. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: a dual-center study. Contrast Media Mol. Imaging 2018.

Iantsen, A., Ferreira, M., Lucia, F., Jaouen, V., Reinhold, C., Bonaffini, P., Alfieri, J., Rovira, R., Masson, I., Robin, P., Mervoyer, A., Rousseau, C., Kridelka, F., Decuypere, M., Lovinfosse, P., Pradier, O., Hustinx, R., Schick, U., Visvikis, D., Hatt, M., 2021. Convolutional neural networks for PET functional volume fully automatic segmentation: development and validation in a multi-center setting. Eur. J. Nucl. Med. Mol. Imaging 1–13.

Iantsen, A., Visvikis, D., Hatt, M., 2021. Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. Lecture Notes in Computer Science (LNCS) Challenges.

Jemaa, S., Fredrickson, J., Carano, R.A., Nielsen, T., de Crespigny, A., Bengtsson, T., 2020. Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks. J. Digit. Imaging 33, 888–894.

Lambin, P., Leijenaar, R.T., Deist, T.M., Peerlings, J., De Jong, E.E., Van Timmeren, J., Sanduleanu, S., Larue, R.T., Even, A.J., Jochems, A., et al., 2017. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 14 (12), 749–762.

Lapuyade-Lahorgue, J., Visvikis, D., Pradier, O., Cheze Le Rest, C., Hatt, M., 2015. SPECTACLE: an automated generalized fuzzy C-means algorithm for tumor delineation in PET. Med. Phys. 42 (10), 5720–5734.

Li, L., Zhao, X., Lu, W., Tan, S., 2019. Deep learning for variational multimodality tumor segmentation in PET/CT. Neurocomputing.

Ma, J., 2021. Cutting-edge 3D medical image segmentation methods in 2020: are happy families all alike? arXiv:2101.00232.

Ma, J., Yang, X., 2021. Combining CNN and hybrid active contours for head and neck tumor segmentation. Lecture Notes in Computer Science (LNCS) Challenges.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9 (1), 1–13.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al., 2020. BIAS: Transparent reporting of biomedical image analysis challenges. Med. Image Anal. 101796.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024.

Moe, Y.M., Groendahl, A.R., Mulstad, M., Tomic, O., Indahl, U., Dale, E., Malinen, E., Futsaether, C.M., 2019. Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. Medical Imaging with Deep Learning.

Naser, M., van Dijk, L., He, R., Wahid, K., Fuller, C., 2021. Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images. Lecture Notes in Computer Science (LNCS) Challenges.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., et al., 2021. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J. Med. Internet Res. 23 (7), e26151.

Parkin, D.M., Bray, F., Ferlay, J., Pisani, P., 2005. Global cancer statistics, 2002. CA Cancer J. Clin. 55 (2), 74–108.

Rao, C., Pai, S., Hadzic, I., Zhovannik, I., Bontempi, D., Dekker, A., Teuwen, J., Traverso, A., 2021. Oropharyngeal tumour segmentation using ensemble 3D PET-CT fusion networks for the HECKTOR challenge. Lecture Notes in Computer Science (LNCS) Challenges.

Ren, J., Eriksen, J.G., Nijkamp, J., Korreman, S.S., 2021. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol. 1–8.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Song, Q., Bai, J., Han, D., Bhatia, S., Sun, W., Rockey, W., Bayouth, J.E., Buatti, J.M., Wu, X., 2013. Optimal co-segmentation of tumor in PET-CT images with context information. IEEE Trans. Med. Imaging 32 (9), 1685–1697.

Vallieres, M., Kay-Rivest, E., Perrin, L.J., Liem, X., Furstoss, C., Aerts, H.J., Khaouam, N., Nguyen-Tan, P.F., Wang, C.-S., Sultanem, K., et al., 2017. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci. Rep. 7 (1), 1–14.

Wahl, R.L., Jacene, H., Kasamon, Y., Lodge, M.A., 2009. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J. Nucl. Med. 50 (Suppl 1), 122S–150S.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.
Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11 (1), 1–15.
Wu, X., Bi, L., Fulham, M., Kim, J., 2020. Unsupervised positron emission tomography tumor segmentation via GAN based adversarial auto-encoder. In: 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV). IEEE, pp. 448–453.
Xie, J., Peng, Y., 2021. The head and neck tumor segmentation using nnU-Net with spatial and channel 'squeeze & excitation' blocks. Lecture Notes in Computer Science (LNCS) Challenges.
Yousefirizi, F., Rahmim, A., 2021. GAN-based bi-modal segmentation using Mumford-Shah loss: application to head and neck tumors in PET-CT images. Lecture Notes in Computer Science (LNCS) Challenges.
Yu, H., Caldwell, C., Mah, K., Poon, I., Balogh, J., MacKenzie, R., Khaouam, N., Tirona, R., 2009. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. Int. J. Radiat. Oncol. Biol. Phys. 75 (2), 618–625.
Yuan, Y., 2021. Automatic head and neck tumor segmentation in PET/CT with scale attention network. Lecture Notes in Computer Science (LNCS) Challenges.
Zhao, X., Li, L., Lu, W., Tan, S., 2018. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. Phys. Med. Biol. 64 (1), 015011.
Zhong, Z., Kim, Y., Zhou, L., Plichta, K., Allen, B., Buatti, J., Wu, X., 2018. 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 228–231.
Zhou, T., Ruan, S., Canu, S., 2019. A review: deep learning for medical image segmentation using multi-modality fusion. Array 100004.
Zhu, S., Dai, Z., Ning, W., 2021. Two-stage approach for segmenting gross tumor volume in head and neck cancer with CT and PET imaging. Lecture Notes in Computer Science (LNCS) Challenges.