Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT

Vincent Andrearczyk[1], Valentin Oreiller[1,2], Moamen Abobakr[3], Azadeh Akhavanallaf[4], Panagiotis Balermpas[5], Sarah Boughdad[2], Leo Capriotti[6], Joel Castelli[7,8,9], Catherine Cheze Le Rest[10,11], Pierre Decazes[6], Ricardo Correia[2], Dina El-Habashy[3], Hesham Elhalawani[12], Clifton D. Fuller[3], Mario Jreige[2], Yomna Khamis[3], Agustina La Greca[5], Abdallah Mohamed[3], Mohamed Naser[3], John O. Prior[2], Su Ruan[6], Stephanie Tanadini-Lang[5], Olena Tankyevych[10,11], Yazdan Salimi[4], Martin Vallières[13], Pierre Vera[6], Dimitris Visvikis[11], Kareem Wahid[3], Habib Zaidi[4], Mathieu Hatt[11], and Adrien Depeursinge[1,2(✉)]

[1] Institute of Informatics, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
adrien.depeursinge@hevs.ch
[2] Department of Nuclear Medicine and Molecular Imaging, Lausanne University Hospital (CHUV), Rue du Bugnon 46, 1011 Lausanne, Switzerland
[3] The University of Texas MD Anderson Cancer Center, Houston, USA
[4] Geneva University Hospital, Geneva, Switzerland
[5] University Hospital Zürich, Zurich, Switzerland
[6] Center Henri Becquerel, LITIS Laboratory, University of Rouen Normandy, Rouen, France
[7] Radiotherapy Department, Cancer Institute Eugène Marquis, Rennes, France
[8] INSERM, U1099, Rennes, France
[9] University of Rennes 1, LTSI, Rennes, France
[10] Centre Hospitalier Universitaire de Poitiers (CHUP), Poitiers, France
[11] LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France
[12] Cleveland Clinic Foundation, Department of Radiation Oncology, Cleveland, OH, USA
[13] Department of Computer Science, Université de Sherbrooke, Sherbrooke, QC, Canada

Abstract. This paper presents an overview of the third edition of the HEad and neCK TumOR segmentation and outcome prediction (HECKTOR) challenge, organized as a satellite event of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022. The challenge comprises two tasks related to the automatic analysis of FDG-PET/CT images for patients with Head and Neck cancer (H&N), focusing on the oropharynx region. *Task 1* is

V. Andrearczyk and V. Oreiller—Equal contribution.
M. Hatt and A. Depeursinge—Equal contribution.

V. Andrearczyk et al. (Eds.): HECKTOR 2022, LNCS 13626, pp. 1–30, 2023.
https://doi.org/10.1007/978-3-031-27420-6_1

the fully automatic segmentation of H&N primary Gross Tumor Volume (GTVp) and metastatic lymph nodes (GTVn) from FDG-PET/CT images. *Task 2* is the fully automatic prediction of Recurrence-Free Survival (RFS) from the same FDG-PET/CT and clinical data. The data were collected from nine centers for a total of 883 cases consisting of FDG-PET/CT images and clinical information, split into 524 training and 359 test cases. The best methods obtained an aggregated Dice Similarity Coefficient (DSC_{agg}) of 0.788 in Task 1, and a Concordance index (C-index) of 0.682 in Task 2.

Keywords: Challenge · Medical imaging · Head and neck cancer · Segmentation · Radiomics · Deep learning · Machine learning

1 Introduction: Research Context

Automatic analysis of multimodal images using machine/deep learning pipelines is of increasing interest. In particular, in the context of oncology, the automation of tumors and lymph nodes delineation can be used for diagnostic tasks (tumor detection), automated staging and quantitative assessment (e.g. lesion volume and total lesion glycolysis), as well as radiotherapy treatment planning and fully automated outcome prediction. This automation presents multiple advantages over manual contouring (faster, more robust and reproducible). Concerning patient-level outcome prediction, multimodal image analysis with machine learning can be used for predictive/prognostic modeling (e.g. response to therapy, prediction of recurrence and overall survival) where image-derived information can be combined with clinical data. These models can be exploited as decision-support tools to improve and personalize patient management.

In this context, the HEad and neCK TumOR segmentation and outcome prediction (HECKTOR) challenge was created in 2020 as a satellite event of MICCAI, with a focus on Head and Neck (H&N) cancer and the use of Positron Tomography Emission / Computed Tomography (PET/CT) images. The first edition of the challenge included a single task, dedicated to the automatic delineation of the primary tumor in combined PET/CT images [6,33]. The second edition (2021) added a second task dedicated to the prediction of Progression-Free Survival (PFS), as well as additional cases from new clinical centers [3]. For the present 2022 (third) edition, the dataset was further expanded (from 425 cases/6 centers to 883 cases/9 centers), and the tasks were updated. For this 2022 edition, Task 1 included the detection and delineation of both the primary tumors and lymph nodes from entire images (no bounding box of the oropharyngeal region was provided as opposed to previous editions), thus achieving a fully automatic segmentation of all pathological targets in the H&N region. Detecting and segmenting both the primary tumor and lymph node volumes opens the avenue to automated TN staging, as well as H&N prognostic radiomics modeling based not only on primary Gross Tumor Volumes (GTVp), but also metastatic lymph nodes (GTVn). The endpoint for Task 2, previously PFS, was change. In

the new edition, Recurrence-Free Survival (RFS) was used, and we focused on automatic prediction (no reference contours were provided for the test cases).

While HECKTOR was one of the first challenges to address tumor segmentation in PET/CT images, other challenges are being organized on this topic. In particular, the AutoPET challenge was organized for the first time in 2022 at MICCAI[1]. The objective of the AutoPET challenge was tumor lesion detection and segmentation in whole-body PET/CT [12]. Based on PET images only, a first challenge on tumor segmentation was previously proposed at MICCAI 2016 [16]. The dataset included both simulated and clinical images. Besides challenges, reviews of general automatic tumor segmentation can be found in [17,37].

Whereas an exponentially increasing number of studies were published on oncological outcome prediction based on PET/CT radiomics [18], challenges addressing this type of task are far less popular than the ones focusing on segmentation tasks. Overall, large-scale validation of both tumor segmentation and radiomics based on PET/CT remains insufficiently addressed, highlighting the importance of this third edition of the HECKTOR challenge.

The paper is organized as follows. Section 2 describes the dataset used for each task. Details concerning evaluation metrics, participation and participants' approaches are detailed in Sects. 3 and 4 for Tasks 1 and 2, respectively. The main findings of the challenges are discussed in Sect. 5 while the conclusions of this 2022 edition are summarized in Sect. 6. Appendix 1 contains additional general information and Appendix 2 details PET/CT image acquisitions.

2 Dataset

2.1 Mission of the Challenge

Biomedical Application
The participating algorithms target the following fields of application: diagnosis, prognosis and research. The participating teams' algorithms were designed for either or both image segmentation (i.e., classifying voxels as either primary tumor, metastatic lymph node or background) and RFS prediction (i.e., ranking patients according to a predicted risk of recurrence). The main clinical motivations for these tasks are introduced in Sect. 1.

Cohorts. As suggested in [28], we refer to the patients from whom the image data were acquired as the challenge cohort. The target cohort[2] comprises patients received for initial staging of H&N cancer.

The clinical goals are two-fold; the automatically segmented regions can be used as a basis for (i) treatment planning in radiotherapy, (ii) further investigations to predict clinical outcomes such as overall survival, disease-free survival,

[1] https://autopet.grand-challenge.org/, as of November 2022.

[2] The target cohort refers to the subjects from whom the data would be acquired in the final biomedical application. It is mentioned for additional information as suggested in BIAS [28], although all data provided for the challenge are part of the challenge cohort.

response to therapy or tumor aggressiveness. The RFS outcome prediction task does not necessarily have to rely on the output of the segmentation task. In the former case (i), the regions will need to be further refined or extended for optimal dose delivery and control. The challenge cohort[3] includes patients with histologically proven H&N cancer who underwent radiotherapy treatment planning. The data were acquired from nine centers (seven for the training, three for the test, including one center present in both sets) with variations in the scanner manufacturers and acquisition protocols. The data contain PET and CT imaging modalities as well as clinical information including center, age, gender, weight, tobacco and alcohol consumption, performance status, HPV status, and treatment (radiotherapy only or additional chemotherapy and/or surgery). A detailed description of the annotations is reported in Sect. 2.2.

Target Entity. The region from which the image data were acquired (called data origin), varied from the head region only to the whole body, and may vary across modalities. Unlike in previous editions [3,33], we provided the data as acquired, without providing automatic bounding-boxes locating the oropharynx regions [4]. The predictions were evaluated on the entire domain of the CT images.

2.2 Challenge Dataset

Data Source
The data were acquired from nine centers as detailed in Table 1. It consists of FDG-PET/CT images of patients with H&N cancer located in the oropharynx region. The devices and imaging protocols used to acquire the data are described in Table 2. Additional information about image acquisition can be found in Appendix 2.

Training and Test Case Characteristics
The training data comprise 524 cases from seven centers (CHUM, CHUS, CHUP, CHUV, HGJ, HMR and MDA). Only patients with complete responses (i.e. disappearance of all signs of local, regional and distant lesions) after treatment are used for Task 2, i.e. 488 cases. The test data contain 359 cases from two other centers CHB and USZ, and from MDA also present in the training set. Similarly, only patients with complete responses after treatment are used for Task 2, i.e. 339 cases. Examples of PET/CT images of each center are shown in Fig. 1. Each case includes aligned CT and PET images, a mask with values 1 for GTVp, 2 for GTVn, and 0 for background (for the training cases) in the Neuroimaging Informatics Technology Initiative (NIfTI) format, as well as patient information (e.g. age, gender) and center.

Participants who wanted to use additional external data for training were asked to also report results using only the HECKTOR data and discuss differences in the results, but none used external data in this edition.

[3] The challenge cohort refers to the subjects from whom the challenge data were acquired.

Table 1. List of the hospital centers in Canada (CA), United States (US), Switzerland (CH) and France (FR) and number of cases, with a total of 524 training and 359 test cases (not all used for task 2, as specified in the rightmost column).

Center	Split	# cases	
		Task 1	Task 2
CHUM: Centre Hospitalier de l'Université de Montréal, Montréal, CA	Train	56	56
CHUS: Centre Hospitalier Universitaire de Sherbooke, Sherbrooke, CA	Train	72	72
HGJ: Hôpital Général Juif, Montréal, CA	Train	55	55
HMR: Hôpital Maisonneuve-Rosemont, Montréal, CA	Train	18	18
CHUP: Centre Hospitalier Universitaire Poitiers, FR	Train	72	44
CHUV: Centre Hospitalier Universitaire Vaudois, CH	Train	53	46
MDA: MD Anderson Cancer Center, US	Train	198	197
Total	Train	524	488
CHB: Centre Henri Becquerel, FR	Test	58	38
MDA: MD Anderson Cancer Center, US	Test	200	200
USZ: UniversitätsSpital Zürich, SW	Test	101	101
Total	Test	359	339

Table 2. List of scanners used in the nine centers. Discovery scanners are from GE Healthcare, Biograph from Siemens, and Gemini from Phillips.

	HGJ	CHUS	HMR	CHUM	CHUV	CHUP	MDA	USZ	CHB	Total
Discovery STE			18	56			133	52		258
Discovery RX							128	24		152
Discovery ST	55						84			139
Biograph 40						72	2			74
Gemini GXL 16		72					1			73
Discovery 690					53		10	8		71
Discovery 710							11		58	69
Discovery HR							2	12		14
Discovery LS							8	3		11
Biograph 64							7			7
Discovery MI							5			5
Biograph 6							1	1		2
Other							2			2
Discovery IQ							1			1
Discovery 600							1			1
Biograph 128								1		1
Biograph 20							1			1
Biograph 16							1			1

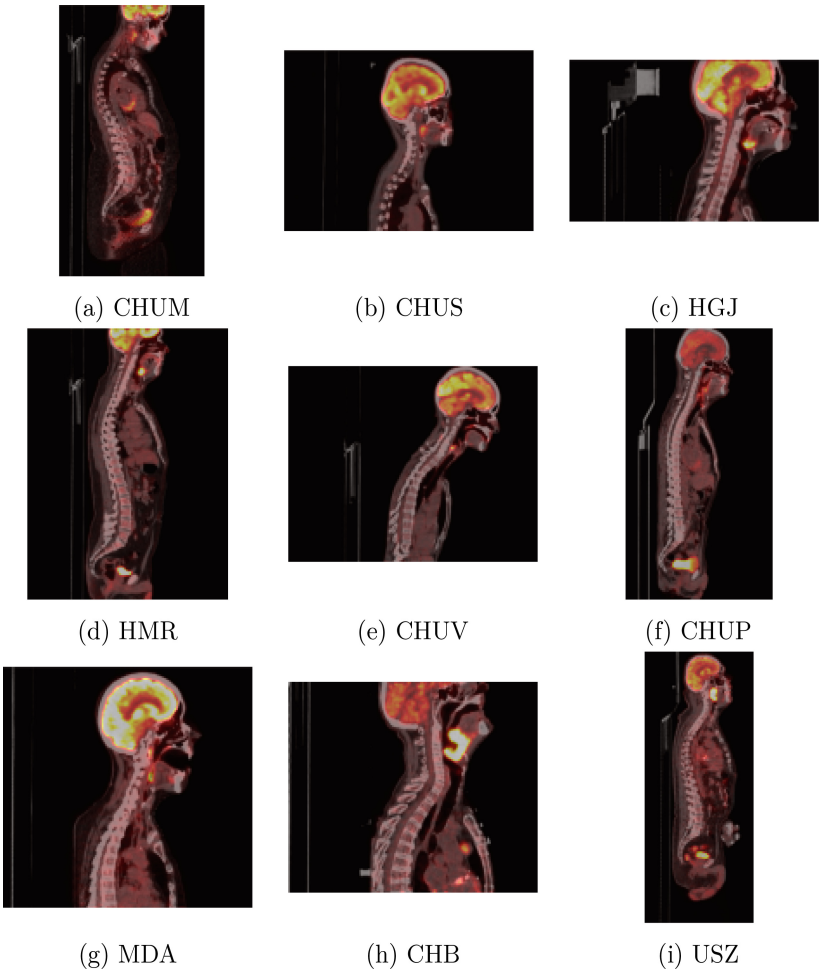


Fig. 1. Case examples of 2D sagittal slices of fused PET/CT images from each of the nine centers, showing the variety of fields of view. The CT (grayscale) window in Hounsfield unit is $[-140, 260]$ and the PET window in SUV is $[0, 12]$, represented in a "hot" colormap.

Task 1 - Ground Truth
Original annotations were performed differently depending on the centers.

– Training set CHUV, CHUS, HGJ, HMR: Contours defining the GTVp and GTVn were drawn by an expert radiation oncologist in a radiotherapy treatment planning system. 40% (80 cases) of the training radiotherapy contours were directly drawn on the CT of the PET/CT scan and thereafter used for treatment planning. The remaining 60% (121) of the training radiotherapy contours were drawn on a different CT scan dedicated to treatment plan-

ning and were then registered to the FDG-PET/CT scan reference frame using intensity-based free-form deformable registration with the software MIM (MIM software Inc., Cleveland, OH). For the training cases, the original number of annotators is unknown.
- Training set CHUV: The GTVp and GTVn were manually drawn on each FDG-PET/CT by a single expert radiation oncologist.
- Training set CHUP: the metabolic volume of primary tumors was automatically determined with the PET segmentation algorithm Fuzzy Locally Adaptive Bayesian (FLAB) (Hatt et al. 2009) and was then edited and corrected manually by a single expert based on the CT image, for example, to correct cases where the PET-defined delineation included air or non-tumoral tissues in the corresponding CT.
- Training and test sets MDA: Contours available from radiotherapy planning (contoured on the CT image, using a co-registered PET as the secondary image to help physicians visualize the tumor) were refined according to the guidelines mentioned below.
- Test set USZ: The primary tumor was separately segmented in the CT and PET images. The CT segmentation was performed manually. In all cases, two radiation oncologists, both having more than 10 years of experience, were involved in the process. Contours were later post-processed for the presence of metal artifacts to exclude non-tumor-related effects. If a certain tumor slice was affected by any artifacts, the entire tumor contour was erased from that slice. Tumors with more than 50% of volume not suitable for the analysis were not included in the study. Additionally, the voxels outside of soft tissue Hounsfield unit (HU) range (20 HU to 180 HU) were discarded. The tumor in the PET image was auto-segmented using a gradient-based method implemented in MIMVISTA (MIM Software Inc., Cleveland, OH).
- Test set CHB: For each patient, the GTVp and GTVn were manually drawn by using the software PET VCAR (GE Healthcare) on each FDG-PET/CT by senior nuclear medicine physicians using adaptive thresholding with visual control using merged PET and CT information.

Quality controls were performed by experts on all the datasets (training and test) to ensure consistency in ground-truth contours definition. The experts re-annotated them, when necessary, to the real tumoral volume (often smaller than volumes delineated for radiotherapy). A shared cloud environment (MIM Cloud Software Inc.) was used to centralize the contouring task and homogenize annotation software. For cases without original GTVp or GTVn contours for radiotherapy, the experts annotated the cases using PET/CT fusion and N staging information. A guideline was developed by the board of experts for this quality control, reported in the following. Cases with misregistrations between PET and CT were excluded. The annotation guidelines are reported in the following.

Guidelines for primary tumor annotation in PET/CT images. The guidelines were provided to the participants during the challenge.

Oropharyngeal lesions are contoured on PET/CT using information from PET and unenhanced CT acquisitions. The contouring includes the entire edges of the morphologic anomaly as depicted on unenhanced CT (mainly visualized as a mass effect) and the corresponding hypermetabolic volume, using PET acquisition, unenhanced CT and PET/CT fusion visualizations based on automatic co-registration. The contouring excludes the hypermetabolic activity projecting outside the physical limits of the lesion (for example in the lumen of the airway or on the bony structures with no morphologic evidence of local invasion). Standardized nomenclature per AAPM TG-263: GTVp. Special situations: Check clinical nodal category to make sure you excluded nearby FDG-avid and/or enlarged lymph nodes (e.g. submandibular, high level II, and retropharyngeal) In case of tonsillar fossa or base of tongue fullness/enlargement without corresponding FDG avidity, please review the clinical datasheet to rule out pre-radiation tonsillectomy or extensive biopsy. If so, this case should be excluded.

Guidelines for nodal metastases tumor annotation in PET/CT images.

Lymph nodes are contoured on PET/CT using information from PET and unenhanced CT acquisitions. The contouring includes the entire edges of the morphologic lymphadenopathy as depicted on unenhanced CT and the corresponding hypermetabolic volume, using PET acquisition, unenhanced CT and PET/CT fusion visualizations based on automatic co-registration for all cervical lymph node levels. Standardized nomenclature for lymph node ROI: GTVn. The contouring excludes the hypermetabolic activity projecting outside the physical limits of the lesion (for example on the bordering bony, muscular or vascular structures).

Task 2 - Ground Truth

The patient outcome ground truths for the prediction task were collected in patients' records as registered by clinicians during patient follow-ups. These include locoregional failures and distant metastases. The time t=0 is set to the end date of radiotherapy treatment.

Data Preprocessing Methods

No preprocessing was performed on the images to reflect the diversity of clinical data and to leave full flexibility to the participants. However, we provided various snippets of code to load, crop and resample the data, as well as to evaluate the results on our GitHub repository[4]. This code was provided as a suggestion to help the participants and to maximize transparency (for the evaluation part). The participants were free to use other methods.

Sources of Errors. A source of error originates from the degree of subjectivity in the annotations of the experts [13,33]. Another source of error is the dif-

[4] https://github.com/voreille/hecktor, as of November 2022.

ference in the re-annotation between the centers used in HECKTOR 2020 and the one added in HECKTOR 2021/2022. In HECKTOR 2020, the re-annotation was checked by only one expert while for HECKTOR 2021/2022 three experts participated in the re-annotation. Moreover, the softwares used were different.

Finally, another source of error comes from the lack of CT images with a contrast agent for a more accurate delineation of the primary tumor.

Institutional Review Boards
Institutional Review Boards (IRB) of all participating institutions permitted the use of images and clinical data, either fully anonymized or coded, from all cases for research purposes only. More details are provided in Appendix 1.

3 Task 1: Segmentation

3.1 Methods: Reporting of Challenge Design

A summary of the information on the challenge organization is provided in Appendix 1, following the BIAS recommendations.

Assessment Aim. The assessment aim for the segmentation task is to evaluate the feasibility of fully automatic GTVp and GTVn segmentation for H&N cancers via the identification of the most accurate segmentation algorithm.

Assessment Method. The performance is measured by the aggregated Dice Similarity Coefficient (DSC_{agg}) between prediction and manual expert annotations. The DSC_{agg} is computed as followed.

$$\mathrm{DSC}_{\mathrm{agg}} = \frac{2\sum_i |A_i \cap B_i|}{\sum_i |A_i| + |B_i|}, \tag{1}$$

with A_i and B_i respectively the ground truth and predicted segmentation for image i, where i spans the entire test set. This metric was employed in [5].

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 and ranking of segmentation algorithms and particularly tumor segmentation tasks. However, the DSC can be problematic, for instance, for cases without ground truth volume, where a single false negative results in a DSC of 0. GTVn is not present in all images and, if present, there can be more than one volume. The predictions can also include zero, one or more volumes. The proposed DSC_{agg} is well-suited to evaluate this type of task. DSC_{agg} will be computed separately for GTVp and GTVn to account for the smaller number of GTVn. The goal is to identify segmentation methods that perform well on the two types of GTV. A drawback of this metric is that standard deviation (or any statistics) across patients cannot be measured.

3.2 Results: Reporting of Segmentation Task Outcome

Participation. As of September 5, 2022 (submission deadline), the number of registered teams for the challenge (regardless of the tasks) was 121. Each team could submit up to three valid submissions. In order to ensure this limit of submissions, only one participant per team was accepted on grand-challenge and allowed to submit results. By the submission deadline, we had received 67 valid submissions for Task 1, i.e. not accounting for invalid submissions such as format errors. This participation was lower than last year's challenge [3].

In this section, we present the algorithms and results of participants in Task 1 with an accepted paper [1,9,10,21,22,24,25,29,30,32,34–36,38–41,43,45–47,49]. A full list of results can be seen on the leaderboard[5].

Segmentation: Summary of Participants' Methods. This section summarizes the approaches proposed by all teams for the automatic segmentation of the primary tumor and metastatic lymph nodes (Task 1). The paragraphs are ordered according to the official ranking, starting with the winners of Task 1. Only a brief description of the methods is provided, highlighting the main particularity, without listing the most commonly used training procedures and parameters such as ensembling, losses etc.

In [32], Myronenko et al. used a SegResNet [31] (a 3D U-Net-like architecture with additional auto-encoder and deep supervision) relying on the MONAI[6] platform, adapted to the specificity of the task (e.g. PET/CT, cropping) with the Auto3DSeg[7] system to automate the parameter choice. The main parts of the pipeline involve image normalization, tumor region detection (specific to HECKTOR 2022), isotropic re-sampling, 5-fold cross-validation, and model ensembling. The tumor region detection is based on relative anatomical positions. Random 3D crops are used for training, centered on the foreground classes with probabilities of 0.45 for tumor, 0.45 for lymph nodes and 0.1 for background.

In [41], Sun et al. employed a coarse-to-fine approach with a cascade of multiple networks. 1) The head is first located in CT with a 3D U-Net. 2) A coarse segmentation of GTVp and GTVn regions is performed with a nnU-Net on PET/CT. The ground truth for this step is taken as the center of the GTVp and GTVn. The output is a smaller bounding box centered on the region of interest. 3) Fine segmentation on PET-CT in the finer bounding box is carried out by an ensemble of five nnU-Nets and five nnFormers (trained with cross-validation) using a 3D SE-norm U-Net to generate the final segmentations of GTVp and GTVn.

In [22], Jiang et al. employed an off-the-shelf nnU-NET with simple pre- and post-processing rules. For training, images are cropped around the GTVp. A post-processing outlier removal is based on minimum volume requirements and distance between predicted GTVp and GTVn volumes. Interestingly, an

[5] https://hecktor.grand-challenge.org/evaluation/challenge/leaderboard/.
[6] https://github.com/Project-MONAI/MONAI.
[7] https://monai.io/apps/auto3dseg.

integration into a web based platform is proposed for the visualization of the segmentation results, including the segmentation for Organs At Risk (OAR), outside the scope of this challenge.

In [34], Rebaud et al. used a simple nnU-Net-based approach with minor adaptations to the task. In particular, images are resampled to $2 \times 2 \times 2$ mm3, and the training is performed on the entire training set after 5-fold and bagging. Median filtering is used to smooth the resampled masks in the CT resolution.

In [35], Salahuddin et al. proposed a 3D U-Net with a channel-wise attention mechanism, grid-attention gates, carefully designed residual connections and dedicated post-processing to remove outlier volumes on the z-axis. The method is trained using 5-fold cross-validation with extensive data augmentation. Uncommonly, input images are resampled to a non-isotropic $1 \times 1 \times 3$ mm^3 voxel size.

In [45], Wang et al. proposed a base nnU-Net combined with a Transfiner (Vision Transformer, ViT-like model with reduced computation and memory costs) to refine the output, based on the assumption that most segmentation errors occur at the tumor boundaries. The Transfiner treats inputs in a similar manner to a ViT, but uses an octree decomposition of multiple layers of interest to select relevant patches instead of densely patchifying the entire image.

In [46], Wang et al. performed a simple segmentation based on nnU-Net. No region detection is used as pre-processing of the segmentation model. A dense patch-based approach ($128 \times 128 \times 128$) is used with a post-processing based on the distance between GTVp and GTVn to eliminate GTVn volumes that are too far from the GTVp (>150 mm).

In [21], Jain et al. compared several segmentation models: nnU-Net (2D/3D), MNet and SwinU-Net architectures. Images are first resampled to $1 \times 1 \times 3$ mm spacing and then registered altogether using the case CHUM-021 as a reference. Further cropping based on the location of the center of the skull was used for input of all model families.

In [9], Chen et al. built an ensemble of three 3D nnU-Nets trained with different loss functions (Dice + focal loss, Dice + top K loss and cross entropy). The models only take as input the CT images. The PET images are used in a final post-processing step where the U-Nets predictions are penalized based on SUV in the PET.

In [39], Rezaeijo et al. used the following multi-step pipeline. First, an organ localizer module is combined with a 3D U-Net for refined organ segmentation, then a 3D ResU-Net is used to segment GTVp and GTVn. The input of the pipeline is a weighted combination of registered PET and CT images.

In [29], Meng et al. proposed a segmentation network based on a U-Net architecture and a cascaded survival network based on a DenseNet architecture. The two networks are jointly optimized with a segmentation loss and a survival loss. The pipeline jointly learns to predict the survival risk scores of patients and the segmentation masks of tumor regions. It extends the already proposed deep multi-task survival DeepMTS model to a radiomics-enhanced deep multi-task framework.

In [25], Lyu proposes to use a 3D nnU-Net model optimized with the Dice Top-K loss function. An ensemble of the five models obtained from cross-validation is used to produce the GTVp and GTVn segmentation masks.

In [49], Xu et al. applied the nnU-Net framework to the cropped PET/CT images. The PET/CT images are cropped according to the oropharyngeal region which was found relative to the brain detected on the PET images. Two different types of nnU-Nets were used, a "vanilla" nnU-Net and another version that was fine-tuned on the test (referred to as PLL nnU-Net). A combination of Dice and cross-entropy losses was used to train the networks.

In [47], Wang et al. first employed a 2D Retina U-Net [20] to localize the H&N region, followed by a 3D U-Net for the segmentation of GTVp/GTVn.

In [36], Salmanpour et al. trained a Cascade-Net [42] (a cascade of a detection module followed by a segmentation module) on a weighted fusion of PET and CT images.

In [10], Chu et al. used a Swin U-NETR [15] with the encoder pretrained by self-supervision on a large CT dataset. The model is trained with cropped images using the bounding-box extractor provided by the organizers [4].

In [38], Shi et al. used a 3D U-Net-based architecture with inputs of multiple resolutions inputted at different depths in the model. Four resolutions are obtained by randomly cropping the images to a fixed size and resampling them to four dimensions (144^3 and repeatedly halved). The model is trained without a validation set.

In [24], La Greca et al. finetuned two pretrained 3D U-Nets on fused PET-CT images. Both models are pretrained with chest CT images and finetuned for GTVp and GTVn segmentation, respectively. The H&N region is detected semi-automatically (i.e. corrected if necessary) based on the head geometry on the CT image.

In [1], Ahamed et al. proposed to use a 2D ResNet50 pretrained on ImageNet as an encoder in a U-Net-like architecture for slice-wise segmentation, trained without data augmentation. The 3D predictions are obtained on the test set by stacking the 2D predictions.

In [30], Müller et al. performed the localization of the H&N region using an analysis of the PET and CT signals on the z-axis, followed by a simple 3D U-Net for precisely locating the region. Patches were then used in a standard 3D U-Net approach based on the winners of previous editions [19,48], followed by classification for differentiating GTVt and GTVn using Support Vector Machines (SVM).

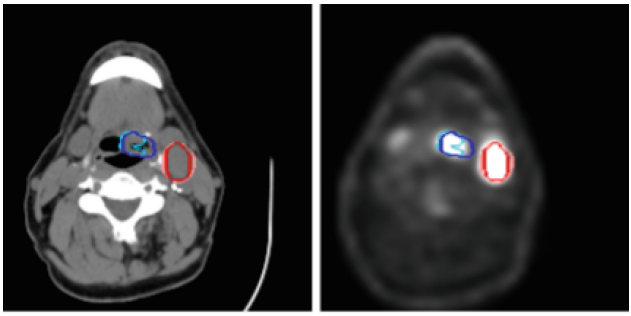
In [43], Thamawita et al. used a cascade of 2D U-Nets (named TriUnet) in order to merge CT-based predictions and PET-based predictions into a single output prediction.

In [40], Srivastava et al. compared three approaches, two based on explicit multi-scale, previously published by the authors, and one based on Swin UNETR [15], a Swin ViT originally designed for brain tumor segmentation. Despite relatively high performance on validation, the generalization to the test data is not optimal with DSC_{agg} values around 0.5.

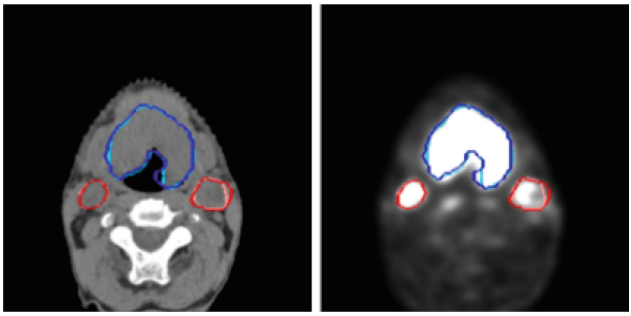
Results. The results are reported in Table 3. The results from the participants range from an average DSC_{agg} of 0.48949 to 0.78802. Myronenko et al. [32] obtained the best overall results with an average DSC_{agg} of 0.78802, respectively 0.80066 on the GTVp and 0.77539 on the GTVn. The best GTVn segmentation was obtained by Sun et al. [41] with a DSC_{agg} of 0.77604. Examples of segmentation results are shown in Fig. 2.

Table 3. Results of Task 1. The best out of three possible submissions is reported for each eligible team. Full list of results available at https://hecktor.grand-challenge.org/evaluation/challenge/leaderboard/.

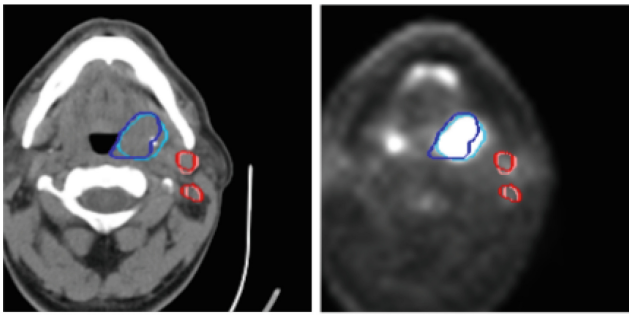
Team	DSC_{agg} GTVp	DSC_{agg} GTVn	mean DSC_{agg}	rank
NVAUTO [32]	**0.80066**	0.77539	**0.78802**	1
SJTU426 [41]	0.77960	**0.77604**	0.77782	2
NeuralRad [22]	0.77485	0.76938	0.77212	3
LITO [34]	0.77700	0.76269	0.76984	4
TheDLab [35]	0.77447	0.75865	0.76656	5
MAIA [45]	0.75738	0.77114	0.76426	6
AIRT [46]	0.76689	0.73392	0.75040	8
AIMers [21]	0.73738	0.73431	0.73584	9
SMIAL [9]	0.68084	0.75098	0.71591	10
Ttest [39]	0.74499	0.68618	0.71559	11
BDAV_USYD [29]	0.76136	0.65927	0.71031	12
junma [25]	0.70906	0.69948	0.70427	13
RokieLab [49]	0.70131	0.70100	0.70115	14
LMU [47]	0.74460	0.65610	0.70035	15
TECVICO Corp [36]	0.74586	0.65069	0.69827	16
RT_UMCG [10]	0.73741	0.65059	0.69400	17
HPCAS [38]	0.69786	0.66730	0.68258	18
ALaGreca [24]	0.72329	0.61341	0.66835	19
Qurit [1]	0.69553	0.57343	0.63448	20
VokCow [30]	0.59424	0.54988	0.57206	21
MLC [43]	0.46587	0.53574	0.50080	22
M&H_lab_NU [40]	0.51342	0.46557	0.48949	23
Average	0.72351	0.68682	0.70517	



(a) MDA-203



(b) CHB-001



(c) USZ-010

Fig. 2. Examples of results of the winning team (NVAUTO [32]). The automatic segmentation results (light) and ground truth annotations (dark) are displayed on an overlay of 2D slices of CT (left) images and PET (right). GTVn is in red and GTVp in blue. CT are clipped between [–140,260] HU and PET images are between [0,5] SUV.

4 Task 2: Outcome Prediction

The second task of the challenge is the prediction of patient outcome, namely RFS.

4.1 Methods: Reporting of Challenge Design

Due to the connection between the two tasks, this second task was carried out on the same dataset as the first one, exploiting both the available clinical information and the multimodal FDG-PET/CT images. Some patients, however, were not used in the second task (see Table 1) because they did not have a complete response to treatment, which is a pre-requisite for the definition of RFS.

The clinical factors included center, age, gender, weight, tobacco and alcohol consumption, performance status, HPV status, and treatment (radiotherapy only or additional chemotherapy and/or surgery). The information regarding tobacco and alcohol consumption, performance status, HPV status and treatment was available only for some patients. The weight was missing in six training and two test cases, and was estimated to 75 kg to compute the Standard Uptake Values (SUV).

Assessment Aim. The chosen clinical endpoint to predict was RFS, i.e. the time t to reappearance of a lesion or to appearance of new lesions (local, regional or distant), censoring deaths. Only patients with complete responses were considered, and death was censored. In the training set, participants were provided with the survival endpoint to predict, censoring and time-to-event between treatment and event (in days). $t = 0$ was defined as the last day of radiotherapy.

Assessment Method. Challengers were asked to submit a CSV file containing the test patient IDs with the outputs of the model as a predicted risk score anti-concordant with the RFS in days. The performance of the predicted scores was evaluated using the Concordance index (C-index) [14] on the test data. The C-index quantifies the model's ability to provide an accurate ranking of the survival times based on the computed individual risk scores, generalizing the area under the ROC curve. It can account for censored data and represents the global assessment of the model discrimination power. The final ranking was based on the best C-index value obtained on the test set out of the maximum of three submissions per team. The C-index computation is based on the implementation in the Lifelines library [11].

4.2 Results: Reporting of Challenge Outcome

Participation. As mentioned for the first task, the number of registered teams for the challenge (regardless of the tasks) was 121. Each team could submit up to three valid submissions. By the submission deadline, we had received 44 valid submissions for Task 2, i.e. not including invalid submissions e.g. due to format errors. All participants of Task 2 also participated in Task 1.

Outcome Prediction: Summary of Participants' Methods. In this section, we describe the algorithms and results of participants in Task 2 [9,25,26,29,30,34–36,43,46,47,49]. A full list of results can be seen on the leaderboard[8].

[8] https://hecktor.grand-challenge.org/evaluation/challenge/leaderboard/.

In [34], Rebaud et al. relied on the Pyradiomics [44] software to extract 93 standard (shape, intensity, textures) radiomics features from the merged GTVp and GTVn delineated volumes (i.e., the result from their task 1 participation) on PET and CT images. In addition to this merged mask, they generated a number of additional masks (re-segmentation with various thresholds, dilation, etc.), which resulted in more than 2400 features per patient. Clinical features were also added, as well as three handcrafted features: the number of tumor masses, the number of lymph nodes, and a binary variable indicating whether the scan was a whole-body scan or included only the H&N region. Each feature was evaluated with the C-index and all pairs of features were also evaluated for their correlation. A novel binary-weighted method was used to assign a binary (-1 / +1) value to each feature, depending on its variation with recurrence time. Finally, the risk was calculated as the mean across all selected feature z-scores weighted by their binary weight. In order to produce a more robust estimate, multiple ensembled models were trained on a random sampling of the training data, also with a randomly selected number of features. A higher number of models led to better performance, and 10^5 models were used on the test set. To evaluate a model on the train set, a two-hundred-fold Monte Carlo cross-validation was used. The ensemble model thus contained three hyperparameters: the number of features randomly drawn for building a model, and the minimum value of C-index and Pearson correlation coefficient threshold to select features among the ones that were randomly drawn. To reduce the risk of overfitting, three bagged models were evaluated in the train and test sets, increasing gradually the number of hyperparameter sets tested, with 10, 100 or 1000 hyperparameter sets, resulting in test C-index values of 0.670, 0.673 and 0.682 respectively.

In [29], Meng et al. proposed an approach similar to the multi-task model trained jointly for both segmentation and prediction task, already proposed in the 2021 edition. The segmentation part is described in Section 3.2. Regarding the outcome prediction task, the model contains a deep learning component, trained on the input PET/CT images, that extracts deep features simultaneously as it generates the segmentation mask. It also contains a standard radiomics component where Pyradiomics features are extracted from the aggregated mask containing the primary tumor and the lymph nodes, as determined by the segmentation part of the pipeline. Finally, clinical variables, deep features and standard radiomics features from the segmentation mask are concatenated in the survival model. The author reported an increased performance in the training set with additional information, with C-index of 0.66, 0.68 and 0.69 relying on clinical variables, standard radiomics and automatic radiomics respectively, whereas the performance of deepMTS only was 0.71, which increased to 0.73 and 0.77 when adding clinical factors then standard radiomics. These three last submissions obtained C-index values of 0.64, 0.65 and 0.68 on the test set. The difference in performance compared with the model ranked 1st is negligible, however, the model is more complex.

In [47], Wang et al. implemented a standard radiomics framework exploiting nnU-Net segmentation, followed by extraction of radiomics features in both PET

and CT images (a single mask containing both GTVp and GTVn, using Pyradiomics), followed by feature selection based on univariate analysis, redundancy through correlation, and finally Cox Proportional Hazard (PH) models building through 5-fold CV for each input (clinical, PET, CT) and a combination of the corresponding risk scores. Regarding the clinical variables, missing values were not imputed but instead coded as a third value. The final model combining risk scores from all three inputs (clinical, PET, CT) obtained a C-index of 0.67 in the test set.

In [26], a standard radiomics approach and a deep learning approach were implemented and compared. For the radiomics approach, features were extracted from a delineated volume containing both the GTVp and the GTVn in a single mask, obtained (in both training and test sets) using the segmentation model of task 1. The authors chose to extract shape and intensity metrics from both modalities, and textural features from the CT component only. Features were extracted with Pyradiomics, using a fixed bin width discretization. Radiomics features were then selected and evaluated in a univariate analysis, as well as using correlation to remove redundant ones. Regarding clinical variables, they were selected based on empirical experience as well as univariate analysis and, in the end, only weight and HPV status were retained. The missing HPV values were not imputed but assigned a third category. A Cox PH model using the selected clinical and hand-crafted radiomics features was then trained. The deep-learning model based on a ResNet and the loss function of DeepSurv [23], trained with data augmentation and oversampling, was implemented to also include the clinical features and the hand-crafted radiomics features selected in the radiomics pipeline. Feature selection, model training and validation (for radiomics and deep learning) were all carried out through a 5-fold CV (based on centers, MDA and HMR centers always in the training). The models evaluated on the test set were obtained by averaging the models obtained in each fold. In the test set, a C-index of 0.668 was obtained using the radiomics approach, whereas the DL model (ensemble of DL and radiomics) obtained 0.646, lower than in the validation (>0.75).

In [49], Xu et al. proposed a standard machine learning approach extracting conventional (volume, SUV, TLG, number of nodes etc.) and radiomics features (SERA package [7]) from both PET and CT modalities. The clinical variables were not exploited in the prognostic models. Cox models were trained using either the conventional features alone, the radiomics alone (with ComBat harmonization based on centers), or the combination of conventional and radiomics features (without harmonization). Other combinations were not studied due to the limited number of test submissions. The best result in the test set was obtained by the conventional model relying on Total Lesion Glycolysis (TLG) and the number of nodes features with a C-index of 0.658, whereas the two more complex models led to lower C-index of 0.645 and 0.648.

In [43], Thambawita et al. proposed at first two approaches, one relying on clinical data only, the other combining clinical variables with basic features from the segmentations (volume and z-extent). In both cases, they used Random

Forest. In a third approach, they combined clinical variables with image data using XGBoost. In addition, they estimated the kidney function of the patients and included it as an additional feature, achieving a C-index of 0.656.

In [30], Müller et al. built upon the winning solution of the past challenge ('Deep Fusion V2') that combines a CNN for extracting deep PET and CT features with a Multi-Layer Perceptron (MLP) trained using a multi-class logistic regression loss (MTLR) for survival tasks. It extends this approach by combining the features (deep, shape and intensity features) extracted from multiple image patches (rather than a single patch) via graph convolution. The resulting embeddings are concatenated with clinical information before the final MLP for MTLR loss training. This multi-patch approach uses as inputs the PET/CT fused images cropped at the segmented tumor centroids. It was trained without data augmentation and was compared to Cox PH and Weibull accelerated failure time models relying on clinical variables and basic tumor descriptors only, the original Deep Fusion V2 models as well as combinations of these. Although the new proposed model performed the best in the validation set (C-index 0.75) it failed in the test set (<0.4). The best result in the test set was obtained with the Weibull model (0.64).

In [25], Lyu et al. proposed a method relying on the AutoGluon[9] framework, which consists in an ensemble of 12 models whose outputs are stacked in several successive layers (here 2 layers). The inputs of the models were standard radiomics features calculated from both the PET and CT images using Pyradiomics. Only three clinical variables were considered (gender, age and chemotherapy). This approach obtained a C-index of 0.639 on the test set.

In [46], Wang et al. trained a ResNet model to predict RFS using, as separate channels, the images (PET only, CT only, or PET/CT), with or without the segmentation mask (output of task 1 using a Retina U-Net) through a 3-fold cross-validation. All investigated combinations led to C-index of 0.64–0.70, with the best model obtained using the PET only (0.70). Its prediction performance on the test set (using an averaging of the three models obtained with 3-fold CV) was 0.635.

In [35], Salahuddin et al. focused mainly on the segmentation task (see Sect. 3.2). Nonetheless, they evaluated the prognostic value of some features extracted from the segmentation masks, namely tumor and lymph node largest volumes and number of lymph nodes through a 5-fold cross-validation. A combination of these three features obtained a C-index of 0.627 on the test set.

In [9], Chen et al. extracted standard radiomics features with Pyradiomics from all the individual lesions predicted by the method of Task 1 (compared to most other challengers who chose to consider the whole segmentation mask). The position (center of mass) of each connected component was also concatenated in the vector of radiomics features. Only clinical variables without missing information were used. Prediction of RFS was achieved by training a multiple-instance neural network in order to handle multiple lesions per patient. Amongst various

[9] https://auto.gluon.ai/stable/index.html.

training strategies (5-fold CV or the entire training set), the best was using the entire training set, reaching a C-index of 0.619 on the test set.

In [36], Salmanpour et al. extracted deep features from the bottleneck of an auto-encoder fed with PET and CT images fused via a weighted technique. These features were selected with mutual information and fed to a random survival forest trained through a 5-fold CV and grid search, obtaining a C-index of 0.59 on the test set.

Results. The results are reported in Table 4.

Table 4. Results of Task 2. The best out of three possible submissions is reported for each eligible team. Full list of results available at https://hecktor.grand-challenge.org/evaluation/challenge/leaderboard/. The predictions of the MLC team were concordant with the time (prediction of days), instead of a risk score. Their C-index results on the leaderboard were, therefore, < 0.5 and they were ranked last on this task. Other teams made this mistake for their first submission, not reported here because we keep only the best results.

Team	C-index	rank
LITO [34]	**0.68152**	1
BDAV_USYD [29]	0.68084	2
AIRT [46]	0.67257	3
RT_UMCG [26]	0.66834	4
RokieLab [49]	0.65817	5
MLC [43]	0.65598	6
VokCow [30]	0.64081	7
junma [25]	0.63896	8
LMU [47]	0.63536	9
TheDLab [35]	0.6305	10
SMIAL [9]	0.61877	11
TECVICO Corp [36]	0.59042	12
Average	0.64769	

The participants' results range from a C-index of 0.59042 to 0.68152, obtained by Rebaud et al. [34].

5 Discussion: Putting the Results into Context

5.1 Outcomes and Findings

Task 1: Automatic Segmentation of GTVp and GTVn
The participation in this task was slightly lower than in the previous edition [3].

This reduction could be partly due to the limit of three submissions instead of five, as well as the increased difficulty of the task arising from (i) not providing bounding-boxes locating the oropharynx region, and (ii) the need to provide segmentation of both GTVp and GTVn. The quality of the methods and their descriptions, however, was improved. Various successful methods were proposed to detect the oropharynx region prior to inputting data into DL models. Without surprise, the best results were obtained with ensembles of 3D U-Nets with careful design choices for pre and post-processing. The use of transformers increased as compared to the previous editions, without clear benefit on the test performance, but achieving very competitive performance. The winner algorithm (NVAUTO [32]) also performed very well on other MICCAI challenges with adaptations to the tasks.

Besides, a surprisingly high performance was obtained in the segmentation of GTVn (DSC_{agg} =0.687 on average and 0.776 for the best), which one may consider more challenging than the primary tumor due to the large variation in location, size and numbers.

Finally, the reported results are not directly comparable with those of 2021 because of the increased complexity (bounding boxes not provided), the different test set (with results highly influenced by tumor sizes), and the different metrics (DSC_{agg} of GTVp and GTVn in 2022 vs average DSC on GTVp in 2021). If we valuate the GTVp DSC in the winner of 2022 to overcome the metric difference, we obtain a DSC of 0.7056 on the 2022 test set, vs 0.7785 in 2021, highlighting the increased complexity of the present edition and the fact that the algorithms are not optimized solely for GTVp segmentation.

Task 2: RFS Prediction

Similarly to Task 1, the participation was lower than last year's challenge which could be due to the increased complexity of the task. In 2021, we observed a majority of deep-learning based pipelines amongst the top results of the prediction task. Four out of the top 5 results relied on deep learning techniques to extract information from PET/CT images and combine it with clinical data to predict PFS, and only one relied on the extraction of engineered radiomics features, combined through ML algorithms. In the current 2022 edition, most of the best results were obtained through the use of standard radiomics features extraction combined with ML modeling, except the second-ranked team (with almost equal performance as the first rank) that relied on a deep learning setting complemented by standard radiomics features. The winner of the 2021 edition ended up ranked 6th in 2022, building on its previously developed purely DL framework. It should also be emphasized that although the number of training and test cases was more than double the number of the previous edition, the overall performance of the predictive model seemed to reach a plateau around C-index 0.7, not better than in 2021. However, a direct comparison between the two editions is challenging since the data is also more heterogeneous (with additional centers being included), the segmentation task was strongly different and more complex, and the clinical endpoint to predict was slightly different (RFS instead of PFS). Of note, most challengers decided to extract features from a

single mask aggregating the primary tumor and the lymph nodes, which may have biased the prognostic value of some of the features, which may have been more relevant when extracted from each lesion type separately.

Finally, no correlation was observed between the results of tasks 1 and 2, i.e. participants who obtained better results in task 2 were not associated with better results in task 1, with a Pearson correlation of 0.065.

5.2 Limitations of the Challenge

Although important efforts were provided by the multidisciplinary consortium to take attention to details in all research aspects that the HECKTOR challenge is addressing, several limitations remain.

The dataset itself presents several limitations. The contours were drawn based on the PET/(unenhanced)CT fusion, although other methods such as MRI with gadolinium or contrast CT are the gold standard to obtain the true contours for radiation oncology. Since the target clinical application is radiomics, however, the precision of the contours is not as important as for radiotherapy planning. Another limitation is due to the variability in the ground truth annotations. Despite the provided guidelines and the quality checks, some heterogeneity in the annotation methods (e.g. in USZ test collection only, removing lesions with metal artifact n and not in other centers, resegmentation in a given HU range) used by the experts and the experts' profiles were observed. Besides our efforts to unify the contours, this led to a remaining significant source of noise in the labelled data used for training.

Concerning Task 1, the segmentation of GTVp and GTVn, one limitation is the segmentation metric which, despite improving the DSC for the task at hand using the aggregated DSC (see Sect. 3.1), is highly biased towards volume sizes. In the future, we could approach the task as a detection problem, using e.g. the refined Dice proposed by Carass et al. [8].

One limitation of Task 2, the prediction of RFS, was the heterogeneity of the patient cohort in terms of HPV status, age groups and other prognostic factors. To mitigate the impact of this limitation, we provided these clinical parameters to the participants, but missing value rates remained high for some variables (e.g. 36% missing values for HPV status). We also worked on the unification of the RFS definition across all centers (see "Assessment aim" in Sect. 4.1), as we realized that even the definition of RFS itself varied across centers or medical specialties. Besides, the treatment information was relatively homogeneous up to the type, but not how exactly the RT was delivered and the combination with chemo (concomitant or subsequent). This is however realistic regarding clinical practice. While the above-mentioned limitations are commonly admitted in the research community on outcome prediction, we think that focusing on clean populations is key to improving the models' performance when one can afford it in terms of sample size.

Finally, this challenge suffers from other known limitations such as the bias of the Dice with respect to tumor size (large tumors obtain higher Dice scores), and limited ranking robustness [27]. The latter two aspects were investigated

for the previous editions [2,33] highlighting a high impact of tumor size on the Dice score and a relatively stable ranking for both segmentation and outcome prediction evaluated with bootstrapping. We expect similar findings for the two tasks of the 2022 edition.

6 Conclusions

This paper presented an overview of the HECKTOR 2022 challenge, dedicated to the automatic analysis of PET/CT images and clinical data of patients with H&N cancer. The tasks proposed in this third edition were (1) Segmentation of primary tumors and metastatic lymph nodes, (2) prediction of patient outcome, namely RFS. The dataset was largely increased in comparison to previous editions, with a total of nine centers and 883 cases. The tasks were also more difficult, in particular with the necessary step of detection of the H&N region prior to further analyses, and the addition of GTVn segmentation to Task 1. Good participation was observed in both tasks, from top research teams across the world proposing a wide variety of methods reported in the 23 quality papers in this volume.

In conclusion, the segmentation results are potentially good enough for clinical use. In future work, we plan to rate the automatic segmentations by experts to assess their quality. Regarding the RFS task, while predictions are better than random, the observed performances suggest that they cannot yet be used clinically to base decisions upon in order to orient treatment options.

Acknowledgments. The organizers thank all the teams for their participation and valuable work. This challenge and the winner prizes were sponsored by Aquilab France, Bioemtech Greece and Siemens Healthineers Switzerland (500€ each, for Task 1, Task 2, and Best Paper). The software used to centralise the annotation and quality control of the GTVp and GTVn regions was MIM (MIM software Inc., Cleveland,OH), which kindly supported the challenge via free licences. This work was also partially supported by the Swiss National Science Foundation (SNSF, grant 205320_179069), the Swiss Personalized Health Network (SPHN, via the IMAGINE and QA4IQI projects) and the RCSO IsNET HECKTOR project.

Appendix 1: Challenge Information

In this appendix, we list additional important information about the challenge as suggested in the BIAS guidelines [28].

Challenge Name
HEad and neCK TumOR segmentation and outcome prediction challenge (HECKTOR) 2022

Organizing Team
The authors of this paper.

Life Cycle Type
A fixed submission deadline was set for the challenge results.

Challenge Venue and Platform
The challenge is associated with MICCAI 2022. Information on the challenge is available on the website, together with the link to download the data, the submission platform and the leaderboard[10].

Participation Policies

(a) Task 1: Algorithms producing fully-automatic segmentation of the test cases were allowed. Task 2: Algorithms producing fully-automatic RFS risk score prediction of the test cases were allowed.
(b) The data used to train algorithms was not restricted. If using external data (private or public), participants were asked to also report results using only the HECKTOR data.
(c) Members of the organizers' institutes could participate in the challenge but were not eligible for awards.
(d) Task 1: The award was 500 euros, sponsored by Aquilab. Task 2: The award was 500 euros, sponsored by Bioemtech. Best paper award: The award was 500 euros, sponsored by Siemens Healthineers Switzerland.
(e) Policy for results announcement: The results were made available on the grand-challenge leaderboard and the best three results of each task were announced publicly. Once participants submitted their results on the test set to the challenge organizers via the challenge website, they were considered fully vested in the challenge, so that their performance results (without identifying the participant unless permission was granted) became part of any presentations, publications, or subsequent analyses derived from the challenge at the discretion of the organizers.
(f) Publication policy: This overview paper was written by the organizing team's members. The participating teams were encouraged to submit a paper describing their method. The participants can publish their results separately elsewhere when citing the overview paper, and (if so) no embargo will be applied.

Submission Method
Submission instructions are available on the website[11] and are reported in the following.

Task 1: Segmentation outputs should be provided as a single label mask per patient (1 for the predicted GTVp, 2 for GTVn, and 0 for the background) in .nii.gz format. The resolution of this mask should be the same as the original CT resolution. The participants should pay attention to saving NIfTI volumes with the correct pixel spacing and origin with respect to the original reference

[10] https://hecktor.grand-challenge.org/.
[11] https://hecktor.grand-challenge.org/Submit/.

frame. The NIfTI files should be named [PatientID].nii.gz, matching the patient names, e.g. CHB-001.nii.gz and placed in a folder. This folder should be zipped before submission. A notebook with a dummy submission example can be found on our github repository[12].

Task 2: Results should be submitted as a CSV file containing the patient ID as "PatientID" and the output of the model (continuous) as "Prediction". An individual output should be anti-concordant with the RFS in days (i.e., the model should output a predicted risk score). If you have a concordant output (e.g. predicted RFS days), you can simply submit your estimate times -1. A notebook with a dummy submission example can be found on our github repository[13].

Participants were allowed three valid submissions per task. The best result was reported in this paper for each task/team.

Challenge Schedule
The schedule of the challenge, including modifications, is reported in the following.

- the release date of the training cases: ~~June 01~~ June 07 2022
- the release date of the test cases: Aug. 01 2022
- the submission date(s): opens Aug. 26 closes ~~Sept. 02~~ Sept. 05 2022 (23:59 UTC-10)
- paper abstract submission deadline: ~~Sept. 02~~ Sept. 05 2022 (23:59 UTC-10)
- full paper submission deadline: Sept. 08 2022 (23:59 UTC-10)
- associated satellite event: Sept. 22 2022

Ethics Approval
Montreal: CHUM, CHUS, HGJ, HMR data (training): The ethics approval was granted by the Research Ethics Committee of McGill University Health Center (Protocol Number: MM-JGH-CR15-50).

CHUV data (training): 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.

CHUP data (training): The fully anonymized data originates from patients who consent to the use of their data for research purposes.

MDA data (training and test): The ethics approval was obtained from the University of Texas MD Anderson Cancer Center Institutional Review Board with protocol number: RCR03-0800.

USZ data (test): The ethics approval was related to the clinical trial NCT01435252 entitled "A Phase II Study In Patients With Advanced Head And Neck Cancer Of Standard Chemoradiation And Add-On Cetuximab".

CHB data (test): The fully anonymized data originates from patients who consent to the use of their data for research purposes.

[12] https://github.com/voreille/hecktor/blob/master/notebooks/example_seg_submission2022.ipynb.
[13] https://github.com/voreille/hecktor/blob/master/notebooks/example_surv_submission2022.ipynb.

Data Usage Agreement
The participants had to fill out and sign an end-user-agreement, available on the grand-challenge platform, in order to be granted access to the data.

Code Availability
The evaluation software was made available on our github page[14]. The participating teams were encouraged to disclose their code.

Conflict of Interest
No conflict of interest applies. Fundings are specified in the acknowledgments. Only the organizers had access to the test cases' ground truth contours.

Appendix 2: Image Acquisition Details

HGJ: For the PET portion of the FDG-PET/CT scan, a median of 584 MBq (range: 368–715) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system (Discovery ST, GE Healthcare). Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 180–420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a span (axial mash) of 5. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52 × 3.52 mm^2 (range: 3.52–4.69). For the CT portion of the FDG-PET/CT scan, an energy of 140 kVp with an exposure of 12 mAs was used. The CT slice thickness resolution was 3.75 mm and the median in-plane resolution was 0.98 × 0.98 mm^2 for all patients.

CHUS: For the PET portion of the FDG-PET/CT scan, a median of 325 MBq (range: 165–517) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system (Gemini GXL 16, Philips). Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 150 s (range: 120–151) per bed position. Attenuation corrected images were reconstructed using a LOR-RAMLA iterative algorithm. The FDG-PET slice thickness resolution was 4 mm and the median in-plane resolution was 4 × 4 mm^2 for all patients. For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 12–140) with a median exposure of 210 mAs (range: 43–250) was used. The median CT slice thickness resolution was 3 mm (range: 2–5) and the median in-plane resolution was 1.17 × 1.17 mm^2 (range: 0.68–1.17).

HMR: For the PET portion of the FDG-PET/CT scan, a median of 475 MBq (range: 227–859) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system (Discovery STE, GE Healthcare). 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

[14] https://github.com/voreille/hecktor.

expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 5 (range: 3–5). The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52×3.52 mm^2 (range: 3.52–5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 120–140) with a median exposure of 11 mAs (range: 5–16) was used. The CT slice thickness resolution was 3.75 mm for all patients and the median in-plane resolution was 0.98×0.98 mm^2 (range: 0.98–1.37).

CHUM: For the PET portion of the FDG-PET/CT scan, a median of 315 MBq (range: 199–3182) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system (Discovery STE, GE Healthcare). 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 (Discovery D690 ToF, GE Healthcare). First, a CT (120 kV, 80 mA, 0.8-s rotation time, slice thickness 3.75 mm) was performed from the base of the skull to the mid-thigh. PET scanning was performed immediately after acquisition of the CT. Images were acquired from the base of the skull to the mid-thigh (3 min/bed position). PET images were reconstructed by using an ordered-subset expectation maximization iterative reconstruction (OSEM) (two iterations, 28 subsets) and an iterative fully 3D (DiscoveryST). CT data were used for attenuation calculation.

CHUP: The acquisition began after 6 h of fasting and 60 ± 5 min after injection of 3 MBq/kg of 18F-FDG (421 ± 98 MBq, range 220–695 MBq), imaged with the PET/CT imaging system (Biograph mCT 40 ToF, Siemens). Non-contrast-enhanced CT images were acquired for attenuation correction (120 kVp, Care Dose® current modulation system) with an in-plane resolution of 0.853×0.853 mm^2 and a 5 mm slice thickness. PET data were acquired using 2.5 min per bed position routine protocol and images were reconstructed using a CT-based attenuation correction and the OSEM-TrueX-TOF algorithm (with time-of-flight and spatial resolution modeling, 3 iterations and 21 subsets, 5 mm 3D Gaussian post-filtering, voxel size $4 \times 4 \times 4$ mm^3).

MDA: For the PET portion of the FDG-PET/CT scan, a median of 401 MBq (range: 327–266) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system (Multiple hybrid

PET/CT scanner devices). Image acquisition of the head and neck was performed using multiple bed positions with a median of 180 s (range: 90–300) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm (2 iterations, 18–24 subsets, 5mm Gaussian filter). The median FDG-PET slice thickness was 3.27 mm (range: 2.99–5) and the median in-plane resolution was $5.46 \times 5.46\,\text{mm}^2$ (range: 2.73×2.73–5.46×5.46). For the CT portion of the FDG-PET/CT scan, a median energy of 120 kVp (range: 100–140) with a median exposure of 185 mAs (range: 10–397) was used. The median CT slice thickness resolution was 3.75mm (range: 2.99–5) and the median in-plane resolution was $0.98 \times 0.98\,\text{mm}^2$ (range: 0.48×0.48–2.734×2.734).

USZ: For PET imaging, an activity of 178–513 MBq was administered intravenously 1h prior to the scan and after the measurement of blood sugar level. Images were acquired with the multiple hybrid PET/CT scanner devices. In the retrospective cohort, 2D or 3D iterative image reconstruction was used, whereas the images of the validation cohort were reconstructed with a 3D algorithm.

CHB: Head and neck PET-CT images were acquired on a GE710 PET/CT device 90 min ($\pm$5 min) after the injection of approximately 3 MBq/kg of FDG. PET and CT acquisition parameters were adapted to the patient's habitus with the patient in the radiotherapy treatment position with a contention mask. For the unenhanced CT portion of the FDG-PET/CT scan, an energy of 120 kVp with an exposure of 25 mAs was used. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm (VPFX, 2 iterations and 23 subsets) 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 $2.73 \times 2.73\,\text{mm}^2$. The CT slice thickness resolution was 2.5 mm and the median in-plane resolution was $0.98 \times 0.98\,\text{mm}^2$ for all patients.

References

1. Ahamed, S., Polson, L., Rahmim, A.: A U-Net convolutional neural network with multiclass Dice loss for automated segmentation of tumors and lymph nodes from head and neck cancer PET/CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
2. Andrearczyk, V., et al.: Automatic head and neck tumor segmentation and outcome prediction relying on FDG-PET/CT images: findings from the second edition of the HECKTOR challenge. Medical Image Analysis (in review)
3. Andrearczyk, V., et al.: Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. HECKTOR 2021. LNCS, vol. 13209, pp. 1–37. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98253-9_1
4. Andrearczyk, V., Oreiller, V., Depeursinge, A.: Oropharynx detection in PET-CT for tumor segmentation. Irish Mach. Vis. Image Process., 109–112 (2020)

5. Andrearczyk, V., Oreiller, V., Jreige, M., Castelli, J., Prior, J.O., Depeursinge, A.: Segmentation and classification of head and neck nodal metastases and primary tumors in PET/CT. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 4731–4735. IEEE (2022)
6. Andrearczyk, V., et al.: Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. In: Andrearczyk, V., Oreiller, V., Depeursinge, A. (eds.) HECKTOR 2020. LNCS, vol. 12603, pp. 1–21. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67194-5_1
7. Ashrafinia, S.: Quantitative nuclear medicine imaging using advanced image reconstruction and radiomics. Ph.D. thesis, The Johns Hopkins University (2019)
8. Carass, A., et al.: Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis. Sci. Rep. **10**(1), 1–19 (2020)
9. Chen, J., Martel, A.: Head and neck tumor segmentation with 3D UNet and survival prediction with multiple instance neural network. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
10. Chu, H., et al.: Swin UNETR for tumor and lymph node delineation of multicentre oropharyngeal cancer patients with PET/CT imaging. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
11. Davidson-Pilon, C.: Lifelines: survival analysis in Python. J. Open Source Softw. **4**(40), 1317 (2019)
12. Gatidis, S., et al.: A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. Sci. Data **9**(1), 1–7 (2022). https://www.nature.com/articles/s41597-022-01718-3
13. Gudi, S., et al.: 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 (2017)
14. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
15. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_22
16. Hatt, M., et al.: The first MICCAI challenge on PET tumor segmentation. Med. Image Anal. **44**, 177–195 (2018)
17. Hatt, M., et al.: Classification and evaluation strategies of auto-segmentation approaches for pet: Report of aapm task group no. 211. Med. Phys. **44**, e1–e42 (2017). https://pubmed.ncbi.nlm.nih.gov/28120467/
18. Hatt, M., et al.: Radiomics in PET/CT: current status and future AI-based evolutions. Seminars Nuclear Med. **51**, 126–133 (2021)
19. Iantsen, A., Visvikis, D., Hatt, M.: Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. In: Andrearczyk, V., Oreiller, V., Depeursinge, A. (eds.) HECKTOR 2020. LNCS, vol. 12603, pp. 37–43. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67194-5_4
20. Jaeger, P.F., et al.: Retina U-NET: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In: Machine Learning for Health Workshop, pp. 171–183. PMLR (2020)

21. Jain, A., et al.: Head and neck primary tumor and lymph node auto-segmentation for PET/CT scans. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
22. Jiang, H., Haimerl, J., Gu, X., Lu, W.: A general web-based platform for automatic delineation of head and neck gross tumor volumes in PET/CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
23. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Med. Res. Methodol. **18**(1), 1–12 (2018)
24. La Greca Saint-Esteven, A., Motisi, L., Balermpas, P., Tanadini-Lang, S.: A fine-tuned 3D U-net for primary tumor and affected lymph nodes segmentation in fused multimodal images of oropharyngeal cancer. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
25. Lyu, Q.: Combining nnUNet and AutoML for automatic head and neck tumor segmentation and recurrence-free survival prediction in PET/CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
26. Ma, B., et al.: Deep learning and radiomics based PET/CT image feature extraction from auto segmented tumor volumes for recurrence-free survival prediction in oropharyngeal cancer patients. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
27. Maier-Hein, L., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. **9**, 5217 (2018). number: 1 Publisher: Nature Publishing Group
28. Maier-Hein, L., et al.: BIAS: transparent reporting of biomedical image analysis challenges. Med. Image Anal. **66**, 101796 (2020)
29. Meng, M., Bi, L., Feng, D., Kim, J.: Radiomics-enhanced deep multi-task learning for outcome prediction in head and neck cancer. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
30. Muller, A.V.J., Mota, J., Goatman, K., Hoogendoorn, C.: Towards tumour graph learning for survival prediction in head & neck cancer patients. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
31. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
32. Myronenko, A., Siddiquee, M.M.R., Yang, D., He, Y., Xu, D.: Automated head and neck tumor segmentation from 3D PET/CT. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
33. Oreiller, V., et al.: Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. Med. Image Anal. **77**, 102336 (2022)
34. Rebaud, L., Escobar, T., Khalid, F., Girum, K., Buvat, I.: Simplicity is all you need: out-of-the-box nnUNet followed by binary-weighted radiomic model for segmentation and outcome prediction in head and neck PET/CT. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
35. Salahuddin, Z., Chen, Y., Zhong, X., Rad, N.M., Woodruff, H., Lambin, P.: HNT-AI: an automatic segmentation framework for head and neck primary tumors and lymph nodes in FDG-PET/CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
36. Salmanpour, M.R., et al.: Deep learning and machine learning techniques for automated PET/CT segmentation and survival prediction in head and neck cancer. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)

37. Savjani, R.R., Lauria, M., Bose, S., Deng, J., Yuan, Y., Andrearczyk, V.: Automated tumor segmentation in radiotherapy. In: Seminars in Radiation Oncology, vol. 32, pp. 319–329. Elsevier (2022)
38. Shi, Y., Zhang, X., Yan, Y.: Stacking feature maps of multi-scaled medical images in U-Net for 3D head and neck tumor segmentation. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
39. Rezaeijo, S.M., Harimi, A., Salmanpour, M.R.: Fusion-based automated segmentation in head and neck cancer via advance deep learning techniques. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
40. Srivastava, A., Jha, D., Aydogan, B., Abazeed, M.E., Bagci, U.: Multi-scale fusion methodologies for head and neck tumor segmentation. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
41. Sun, X., An, C., Wang, L.: A coarse-to-fine ensembling framework for head and neck tumor and lymph segmentation in CT and PET images. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
42. Tang, M., Zhang, Z., Cobzas, D., Jagersand, M., Jaremko, J.L.: Segmentation-by-detection: a cascade network for volumetric medical image segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1356–1359. IEEE (2018)
43. Thambawita, V., Storas, A., Hicks, S., Halvorsen, P., Riegler, M.: LC at HECKTOR 2022: the effect and importance of training data when analyzing cases of head and neck tumors using machine learning. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
44. Van Griethuysen, J.J., et al.: Computational radiomics system to decode the radiographic phenotype. Can. Res. **77**(21), e104–e107 (2017)
45. Wang, A., Bai, T., Jiang, S.: Octree boundary transfiner: efficient transformers for tumor segmentation refinement. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
46. Wang, K., et al.: Recurrence-free survival prediction under the guidance of automatic gross tumor volume segmentation for head and neck cancers. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
47. Wang, Y., et al.: Head and neck cancer localization with Retina Unet for automated segmentation and time-to-event prognosis from PET/CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)
48. Xie, J., Peng, Y.: The head and neck tumor segmentation using nnU-Net with spatial and channel squeeze & excitation blocks. In: Andrearczyk, V., Oreiller, V., Depeursinge, A. (eds.) HECKTOR 2020. LNCS, vol. 12603, pp. 28–36. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67194-5_3
49. Xu, H., Li, Y., Zhao, W., Quellec, G., Lu, L., Hatt, M.: Joint nnU-Net and radiomics approaches for segmentation and prognosis of head and neck cancers with PET/CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2023)