Contents lists available at ScienceDirect

Medical Image Analysis

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

Automatic Head and Neck Tumor Segmentation and Outcome Prediction Relying on FDG-PET/CT Images: Findings from the Second Edition of the HECKTOR Challenge

Vincent Andrearczyk[*a], Valentin Oreiller[*a,b,c], Sarah Boughdad[b], Catherine Cheze Le Rest[d,e], Olena Tankyevych[d,e], Hesham Elhalawani[f], Mario Jreige[b], John O. Prior[b,c], Martin Vallières[g], Dimitris Visvikis[d], Mathieu Hatt[**d], Adrien Depeursinge[**a,b]

[a]*Institute of Informatics, 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]*Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland*
[d]*LaTIM, INSERM, UMR 1101, University Brest, Brest, France*
[e]*Poitiers University Hospital, nuclear medicine, Poitiers, France*
[f]*Cleveland Clinic Foundation, Department of Radiation Oncology, Cleveland, OH, US*
[g]*Department of Computer Science, Université de Sherbrooke, Sherbrooke, Québec, Canada*

ARTICLE INFO

Article history:

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Medical Imaging, Head and Neck Cancer, Automatic Segmentation, Radiomics

ABSTRACT

By focusing on metabolic and morphological tissue properties respectively, FluoroDeoxyGlucose (FDG)-Positron Emission Tomography (PET) and Computed Tomography (CT) modalities include complementary and synergistic information for cancerous lesion delineation and characterization (e.g. for outcome prediction), in addition to usual clinical variables. This is especially true in Head and Neck Cancer (HNC). The goal of the HEad and neCK TumOR segmentation and outcome prediction (HECKTOR) challenge was to develop and compare modern image analysis methods to best extract and leverage this information automatically. We present here the post-analysis of HECKTOR 2nd edition, at the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2021. The scope of the challenge was substantially expanded compared to the first edition, by providing a larger population (adding patients from a new clinical center) and proposing an additional task to the challengers, namely the prediction of Progression-Free Survival (PFS). To this end, the participants were given access to a training set of 224 cases from 5 different centers, each with a pre-treatment FDG-PET/CT scan and clinical variables. Their methods were subsequently evaluated on a held-out test set of 101 cases from two centers. For the segmentation task (Task 1), the ranking was based on a Borda counting of their ranks according to two metrics: mean Dice Similarity Coefficient (DSC) and median Hausdorff Distance at 95^{th} percentile (HD95). For the PFS prediction task, challengers could use the tumor contours provided by experts (Task 3) or rely on their own (Task 2). The ranking was obtained according to the Concordance index (C-index) calculated on the predicted risk scores. A total of 103 teams registered for the challenge, for a total of 448 submissions and 29 papers. The best method in the segmentation task obtained an average DSC of 0.759, and the best predictions of PFS obtained a C-index of 0.717 (without relying on the provided contours) and 0.698 (using the expert contours). An interesting finding was that best PFS predictions were reached by relying on DL approaches (with or without explicit tumor segmentation, 4 out of the 5 best ranked)

[*]These authors contributed equally to this work.
[**]These authors contributed equally to this work.

compared to standard radiomics methods using handcrafted features extracted from delineated tumors, and by exploiting alternative tumor contours (automated and/or larger volumes encompassing surrounding tissues) rather than relying on the expert contours. This second edition of the challenge confirmed the promising performance of fully automated primary tumor delineation in PET/CT images of HNC patients, although there is still a margin for improvement in some difficult cases. For the first time, the prediction of outcome was also addressed and the best methods reached relatively good performance (C-index above 0.7). Both results constitute another step forward toward large-scale outcome prediction studies in HNC.

1. Introduction

Combined FluoroDeoxyGlucose (FDG)-Positron Emission Tomography (PET) and Computed Tomography (CT) are now considered the image modalities of choice for diagnosis, treatment planning, and therapy response evaluation in a number of pathologies, including Head and Neck Cancer (HNC). PET (with FDG or other radiotracers) and CT provide complementary, synergistic and quantitative data, both functional and anatomical. These data, combined with the usual clinical variables (e.g. clinical stage, age, gender, etc.) have been shown to be useful for patient management and to have an impact on cancerous lesion delineation (e.g. for radiotherapy planning purposes) and characterization (e.g. for diagnosis, therapy response evaluation or outcome prediction). However, the PET/CT images remain exploited mostly in a manual and visual fashion in the clinical routine practice, despite the deployment of semi-automated tools in clinical stations. The current efforts of the research community are therefore aimed at the development of tools that are efficient, trustworthy (i.e. interpretable and generalizable) and as automated as possible (without excluding the clinicians from the loop) in order to better exploit the quantitative content of the available images, a methodological approach which is today termed radiomics (Lambin et al., 2012; Gillies et al., 2016). Although the recognition (i.e. detection) and contouring (i.e. segmentation) tasks have long been the focus of specific developments, for example for radiotherapy treatment planning automation (Harrison et al., 2022) or the help in diagnosis and tumor burden quantification, they are also a crucial part of the usual radiomics pipeline. They are exploited after image pre-processing and before the characterization of the detected and delineated tumor volume through handcrafted image features (e.g. shape, intensity, texture) that are subsequently used for modeling (Lambin et al., 2017). In both fields of segmentation and outcome prediction, the rise of approaches entirely or partly based on deep learning (DL) has been particularly striking over the last few years. On the one hand, for the detection and/or segmentation, most of the algorithms proposed in the literature and in the context of medical image challenges are now based on DL, especially the U-Net (Ronneberger et al., 2015) architecture which has been very successful (Savjani et al., 2022; Eisenmann et al., 2022). On the other hand, in the field of outcome prediction (radiomics), DL has not yet entirely supplanted the use of handcrafted features exploited through classical Machine Learning (ML) algorithms. In 2020, the HECKTOR challenge was organized for the first time, in the context of the MICCAI conference. Its focus and its unique task was the automated segmentation of the primary Gross Tumor Volume (GTVp) in combined PET/CT images of HNC patients. All challengers proposed solutions based on variants of U-Net (Andrearczyk et al., 2020b). Because of the success of the challenge for its first edition, it was decided to renew it in 2021. The scope of this second edition has been substantially expanded, with the addition of patients and centers as well as a second task, outcome prediction, which could be addressed independently from the segmentation task. This paper provides a post-challenge

*Corresponding author
e-mail: vincent.andrearczyk@hevs.ch (Vincent Andrearczyk*)

analysis and reports the most relevant findings of HECKTOR 2021. Additionally to the raw presentation of the data, participation and results in (Andrearczyk et al., 2021b), we largely extend the data description and analysis of the results, including ranking stability, statistical tests, ensembles of predictions, comparison with PET thresholding methods, influence of tumor size and SUV on segmentation performance, inter-center performance, analysis of the overfitting resulting from the best of 5 submissions per team, as well as additional baseline results for the outcome prediction.

2. Prior Work

2.1. Related Tumor Segmentation Algorithms

Before the DL era, standard segmentation methods in computer vision and medical imaging included thresholding, region-based (e.g. region growing), watershed and clustering (e.g. K-means) (Foster et al., 2014). Tumor segmentation in PET images is still commonly performed with Standardized Uptake Values (SUV) thresholding in clinical routine, yet it is difficult to fully automatize due to the semi-quantitative nature of SUVs that vary e.g. with the time between the injection and acquisition, the scanner, the reconstruction algorithm, the tumor shape, and individual normal physiological baseline (Wahl et al., 2009).

A major breakthrough in medical image segmentation, at the basis of most current algorithms, came from the adaptation of deep Convolutional Neural Networks (CNN) to the field. U-Net (Ronneberger et al., 2015) and its 3D variants (Çiçek et al., 2016) are designed as fully-convolutional encoder-decoder networks with skip connections to leverage contextual information and precise localization. Based on the success of Squeeze and Excitation networks (SENets) (Hu et al., 2018), a squeeze-and-excitation normalization layer was introduced in a U-Net architecture by Iantsen et al. (Iantsen et al., 2020) in HECKTOR 2020, reaching the best performance for HNC primary tumor segmentation. nnU-Net (”no new net”) (Isensee et al., 2021) has recently become a popular method for various medical image segmentation tasks. Developed to deal with datataset diversities, it automates key design choices for a successful segmentation pipeline using standard 2D or 3D U-Nets.

Vision transformers (ViT) (Dosovitskiy et al., 2020) are self-attention-based architectures inspired by natural language processing transformer models (Vaswani et al., 2017). Images are split into a sequence of patches, combined together with a position embedding and fed to a transformer encoder. One major difference compared with CNNs is the early aggregation of global information, as well as a less constrained architecture that tends to perform better with large amounts of data. A segmentation ViT (U-NeTr) was proposed in (Hatamizadeh et al., 2022), combining the strength of the encoder-decoder and skip connections of the U-Net with the self-attention mechanism of the ViT. This architecture was applied to PET/CT tumor segmentation in (Sobirov et al., 2022). This method did not reach the highest performance of CNNs, yet further developments and adaptations of ViTs may play an important role in PET/CT tumor segmentation and other prediction tasks in the coming years, as suggested by their increasing use in the medical imaging field and recent developments (Liu et al., 2021).

2.2. Medical Image Segmentation Challenges

The interest in medical imaging challenges has grown over the past years. These challenges have enabled a fair comparison of algorithms developed by various research teams across the world on large curated datasets, as opposed to studies performed on non-public data, and/or diverging data splits and evaluation protocols. The number of MICCAI challenges between 2018 and 2022 has increased from 15 to 38 (25 of which feature a segmentation task in 2022), equally supported by a growing participation. A similar trend is observed in other challenges organized independently or at other venues 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). Quality of the data, as well as fairness and appropriateness of the evaluation,

have greatly benefited from initiatives such as challenge design guidelines (Maier-Hein et al., 2018) and Biomedical Image Analysis ChallengeS (BIAS) reporting guidelines (Maier-Hein et al., 2020). The Brain Tumor Segmentation (BraTS) challenge (Menze et al., 2014) is an exemplary successful challenge that has evolved in the past decade, has proposed various tasks and data and has been the opportunity for multiple key technical developments in the field of medical imaging.

Following the first PET tumor segmentation challenge (Hatt et al., 2018), HECKTOR 2020 (Oreiller et al., 2022) was the first challenge addressing the segmentation of cancer-related lesions in PET/CT. It featured a single task of primary tumor segmentation (GTVp) in HNC patients. In 2021, we increased the dataset size and added a second task of patient outcome prediction (Andrearczyk et al., 2021b).

2.3. Related Outcome Prediction Algorithms

Radiomics (Gillies et al., 2016) is the quantitative analysis of radiological and nuclear medicine images with high throughput extraction to obtain a non-invasive diagnosis and prognosis support for precision medicine. Survival radiomics aims at predicting the elapsed time between the diagnosis or the treatment date and some event of interest. Commonly employed event types includes death (e.g. in overall survival) and recurrence (e.g. in Progression Free Survival, PFS). PFS, used in this challenge, considers progression of the disease, local and regional recurrence, distant metastasis and death of any cause as events. Only the end of follow-up is censored. The time is generally considered from the end of treatment to the event.

Proportional hazards models (e.g. Cox proportional hazards model (Cox, 1972)) are often used in survival analyses to predict the hazard risk using using one or more covariates such as clinical variables or semantic/quantitative image biomarkers in radiomics studies (Burke et al., 1997; Vallières et al., 2017). Supported by its success in various machine vision tasks, CNNs have been applied to radiomics, although standard radiomics approaches involving feature extraction from regions of interest followed by ML models remain popular due to the frequently limited amount of data and the low computational requirements. Survival CNN models use survival losses such as the Cox loss (Zhang et al., 2020; Andrearczyk et al., 2021a) to learn relevant features directly from the data.

In HNC cancer, radiomics has been applied to prognosis prediction from CT (Aerts et al., 2014), MRI (Dang et al., 2015) and, recently, PET/CT (Diamant et al., 2019; Vallières et al., 2017; Andrearczyk et al., 2020c) images. A valuable review of HNC radiomics is provided in (Wong et al., 2016). In (Fontaine et al., 2021), it was shown that delineations made for radiotherapy are not well suited for radiomics. Re-delineating radiotherapy contours improved the prediction of radiomics models for the prediction of DFS in HNC cancer. Together with the low inter-observer agreement reported in (Oreiller et al., 2022) (0.61), these results motivated the definition of delineation guidelines (Section 3.2) and the curation of radiotherapy contours for the HECKTOR data proposed in this paper. The fully automatic radiomics task proposed in this challenge (Task 2, without ground truth test contours given to the participants) was motivated by the preliminary experiments in (Fontaine et al., 2021) showing promising DFS prediction results using automatically segmented tumors as Volumes of Interest (VOIs) for the radiomics pipeline, as well as the multi-task CNN proposed in (Andrearczyk et al., 2021a) to guide the prediction of DFS with an auxiliary tumor segmentation task.

2.4. Challenges Addressing Outcome Prediction

Despite the numerous radiomics studies reported in the past 10 years, challenges on patient outcome prediction have been far less popular than segmentation (e.g. lesion, organs), classification (e.g. diagnosis, staging) and registration challenges. The BraTS (Menze et al., 2014) challenge proposed, among others, the task of OS and pseudoprogression prediction of glioblastoma multiforme from brain MRI. The OroPharynx Cancer (OPC) Radiomics Challenge[1] proposed the task of binary prediction of local

[1] `https://www.kaggle.com/competitions/opc-recurrence/overview`, as of May 2023.

Table 1: List of the hospital centers in Canada (CA), Switzerland (CH) and France (FR) and number of cases, with a total of 224 training and 101 test cases.

Center	Split	# cases
HGJ: Hôpital Général Juif, Montréal, CA	Train	55
CHUS: Centre Hospitalier Universitaire de Sherbooke, Sherbrooke, CA	Train	72
HMR: Hôpital Maisonneuve-Rosemont, Montréal, CA	Train	18
CHUM: Centre Hospitalier de l'Université de Montréal, Montréal, CA	Train	56
CHUP: Centre Hospitalier Universitaire Poitiers, FR	Train	23
Total	Train	224
CHUV: Centre Hospitalier Universitaire Vaudois, CH	Test	53
CHUP: Centre Hospitalier Universitaire Poitiers, FR	Test	48
Total	Test	101

recurrence from CT images, evaluated with AUC. The 18-FDG-PET Radiomics Risk Stratifiers in Head and Neck Cancer challenge, at MICCAI 2018[2] proposed the binary task of local tumor control prediction.

3. HECKTOR 2021 Challenge Set-up

3.1. Dataset

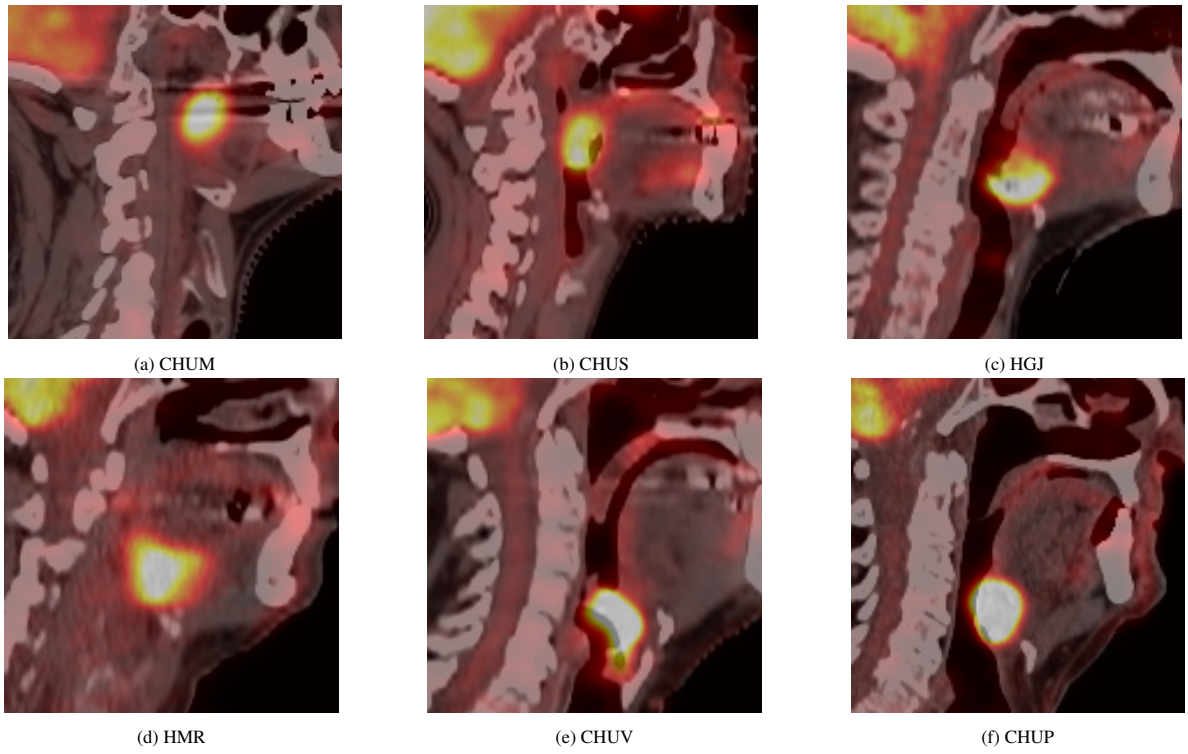
Data source.

Compared to the dataset used in the first edition of the challenge (Oreiller et al., 2022), 71 patients collected in a sixth center (CHUP) were added and distributed between the training (n=23) and the testing (n=48) sets. As a result, the 2021 dataset consisted of patient data and images collected from six centers as detailed in Table 1 (Andrearczyk et al., 2021b). The data consist of PET/CT images of patients with HNC located in the oropharynx region. Inclusion criteria were head and neck cancer patients older than 18 years with primary oropharyngeal lesions of any histologic type, TNM stage and HPV status. Exclusion criteria was unavailable follow-up data. Additional information about the image acquisition is provided in Appendix A.

Training and test case characteristics.

The training data comprise 224 cases from five centers (HGJ, HMR[3], CHUM, CHUS and CHUP). The data from the first four centers originate from (Vallières et al., 2017) which contained 298 cases, among which we selected the cases with oropharynx cancer. The test data contain 101 cases from a sixth center (CHUV n=53) and from CHUP (n=48). Examples of PET/CT images of each center are illustrated in Fig. 1. The HGJ, HMR, CHUM, CHUS cohorts were aleardy used in the training set of HECKTOR 2020 and CHUV cohort was already used in the 2020 test set. The number of events in the training set is 56 out of 224 (25%). In the test set, it is 40 out of 101 events (39.6%). In terms of severity of disease, the distribution of TNM stages in the training set is TNM stages in the test set is I: 4%, II: 6.9%, III: 18.8%, IV: 70.3%. In the training set, it is I: 1.8%, II: 8.5%, III: 12.9%, IV: 76.8%. These variations reflect the clinical reality and require good model generalization to reach high prediction performance. Each case includes aligned PET and CT volumes, a GTVp mask (for the training cases only) in the Neuroimaging Informatics Technology Initiative (NIfTI) format, as well as patient clinical information (age, gender, center, T, N and M stage and clinical staging edition (7th or 8th depending on the center), as well as tobacco and alcohol consumption, performance status, Human PapillomaVirus (HPV) infection status and therapy modalities (i.e. radiotherapy only or chemoradiotherapy). Similarly to the 2020 edition, a bounding-box of size $144 \times 144 \times 144$ mm^3 locating the oropharynx region was also provided. Details of the semi-automatic region detection can be

[2] `https://www.kaggle.com/competitions/pet-radiomics-challenges/overview`, as of May 2023.
[3] For simplicity and consistency, these centers were renamed CHGJ and CHMR during the challenge.



(a) CHUM (b) CHUS (c) HGJ

(d) HMR (e) CHUV (f) CHUP

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

found in (Andrearczyk et al., 2020a). This is an important point to emphasize because it helps the challengers focus their algorithm development on the actual segmentation and outcome prediction, rather than on a recognition/detection step.

3.2. Contours

We describe the delineation of primary tumors for all cases in the dataset, used as ground truth for the segmentation and VOIs for one of the outcome prediction tasks. As described below, some cases had original delineations from radiotherapy which were re-contoured to obtain target contours that are as close as possible to the true tumor boundaries. Since no guidelines were available for contouring true HNC tumoral volumes on PET/unenhanced CT, we defined them to reduce inter-observer variability by reaching a consensus on the best approach to adopt.

Definition 3.1 (GTVp primary tumor delineation guidelines)**.**
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). For more specific situations, the clinical nodal category was verified to ensure the exclusion of nearby FDG-avid and/or enlarged lymph nodes (e.g. submandibular, high level II, and retropharyngeal). In the case of tonsillar fossa or base of tongue fullness/enlargement without corresponding FDG avidity, the clinical datasheet was reviewed to exclude patients with pre-radiation tonsillectomy or extensive biopsy.

The contours for the CHUV center were drawn by an expert radiation oncologist for radiomics purposes (Castelli et al., 2019).

The expert contoured the tumors on fused PET/CT scans. The cases from HGJ, CHUS, HMR, and CHUM centers were originally contoured in the context of radiotherapy (Vallières et al., 2017). All contours were re-delineated for radiomics purposes according to the aforementioned guidelines (Definition 3.1) for HECKTOR 2020 (Oreiller et al., 2022). For the data added to the current HECKTOR 2021 edition (CHUP), the delineations were obtained semi-automatically with a Fuzzy Locally Adaptive Bayesian (FLAB) segmentation (Hatt et al., 2009) applied to the PET image, and subsequently corrected by an expert radiation oncologist based on the corresponding CT information for radiotherapy planning. The re-delineation of true tumoral volume was performed by three experts: one nuclear medicine physician, one radiation oncologist and one who is both radiologist and nuclear medicine physician. The 71 cases were divided between the three experts and each delineation was then cross-checked by all three of them. This re-delineation was performed in a centralized fashion with the MIM software, and the verification of the contours was made possible by the MIM Cloud platform [4]. Overall, although different strategies were used to create the initial contours, these were controlled and corrected to strictly follow the unified guidelines introduced in Definition 3.1.

3.3. Patient Outcome

Clinical information regarding patients was collected in each clinical center along with the PET/CT images. Such clinical variables included age, gender, T, N and M stage and clinical staging (7th or 8th edition depending on the center), as well as tobacco and alcohol consumption, performance status, HPV status and therapy modalities (i.e., radiotherapy only or chemoradiotherapy) when available. Information regarding the chosen endpoint to predict (see the section below), i.e., censoring and time-to-event between PET/CT scan and event (in days) was provided as well (for the training data only).

3.4. Challenge tasks

Task 1 - Primary Tumor Segmentation. The first task is similar to the first edition with a larger dataset, namely the automatic segmentation of GTVp from the PET/CT images. To perform well in this task, algorithms should accurately segment test primary tumors.

Task 2-3 - Outcome Prediction. The new task of outcome prediction was added for this second edition of the challenge. It relied on the same training and testing dataset as the segmentation (Task 1). This design allowed challengers to exploit their segmentation results from Task 1, although this was not mandatory. Outcome prediction was divided into two different tasks and associated submissions and rankings, depending on whether the reference expert contours (the ground truth of Task 1) were exploited by the method or not. Task 2 was therefore defined as predicting outcomes based only on PET/CT images and the available clinical variables. In contrast, Task 3 had the same goal, with the addition of expert contours made available. Exploiting the clinical variables for predicting outcomes was not mandatory. To avoid challengers willing to participate in Task 3 having direct access to the ground truth of Task 1, we requested they encapsulate their methods within Docker containers[5] that we evaluated on the test data. The prediction endpoint was chosen as PFS since this information was available in the clinical data for all included patients. It is a clinically-relevant endpoint that can be leveraged to support decision systems for personalized patient management in such a HNC population. Progression was defined according to RECIST (Response Evaluation Criteria In Solid Tumors): either a size increase of known lesions (i.e., change of T and or N), or the appearance of new lesions (i.e., change of N and/or M). We considered disease-specific death a progression event for patients previously regarded as stable. In the training set, participants were provided with the event info (1 or 0), censoring, and time-to-event between PET/CT scan and event in days. The number of PFS events

[4] `https://mim-cloud.appspot.com/` as of May 2023.
[5] `https://www.docker.com/` as of September 2022.

was 56 in the training set, 40 in the test set. To evaluate and rank challengers, we relied on the Concordance index (C-index), an established metric to quantify predicted risk scores. To perform well in these tasks, algorithms should correctly rank the test cases based on predicted scores of risk of progression.

3.5. Rankings and Assessment Methods

Task 1 - Primary Tumor Segmentation. Participants were given access to the test cases without the ground truth delineations and were asked to submit the results of their algorithms on these cases on the AIcrowd platform[6]. We only accepted binary segmentations in the NIfTI file format.

Results were ranked using the 3D Dice Similarity Coefficient (DSC) and Hausdorff Distance at 95$^{\text{th}}$ percentile (HD95), both computed on images cropped using the provided bounding-boxes (see Section 3.1) in the original CT resolution. The two metrics are defined for set A (ground truth volumes) and set B (predicted volumes) as follows.

$$\mathrm{DSC}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{1}$$

where $|\cdot|$ is the set cardinality and

$$\mathrm{HD95}(A, B) = P_{95}\left\{\sup_{a \in A} \inf_{b \in B} \mathrm{d}(a, b), \sup_{b \in B} \inf_{a \in A} \mathrm{d}(a, b)\right\}, \tag{2}$$

where $\mathrm{d}(a, b)$ is the Euclidean distance between points a and b, sup and inf are the supremum and infimum, respectively. P_{95} is the 95$^{\text{th}}$ percentile. We followed the recommendation of Maier-Hein et al. (2022) to combine an overlap-based metric (i.e. DSC) with a boundary-based metric (i.e. HD95), as the latter compensates bias towards larger structures. HD95 was used instead of HD thanks to its robustness to spatial outliers.

Before the challenge opening, we decided to handle missing predictions by attributing a DSC of 0 and a HD95 of $+\infty$ to them. However, this did not occur 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, recall and F1-score to investigate whether the methods were rather providing a large false positive or false negative rate. The evaluation implementation can be found on our GitHub repository[7] and was provided to the participants to maximize transparency.

The ranking was computed from the average DSC and median HD95 across all cases. Since the HD95 is unbounded, i.e. it is infinity when there is no prediction, we choose the median instead of the mean for aggregation. The two metrics are ranked separately and the final rank is obtained by Borda counting. This ranking method was used first to determine the best submission of each participating team (ranking the one to five submissions), then to obtain the final ranking (across all participants). 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. These errors did not count against their quota of 5 submissions. No immediate feedback was displayed on how their run was performing to avoid iterative overfit.

Tasks 2 and 3 - PFS prediction. Participants were given access to the test cases without the PFS ground truth annotations and were asked to submit the results of their algorithms on these cases on the AIcrowd platform (Task 2) or to encapsulate their algorithms exploiting ground truth expert contours in a Docker that was run by organizers on the test data (Task 3). The expected output of

[6] `https://www.aicrowd.com/challenges/miccai-2021-hecktor`, as of September 2022.
[7] `github.com/voreille/hecktor/tree/hecktor2020/src/evaluation`, as of April 2022.

the algorithm was a CSV file containing the patient ID's along with the predicted risk scores, anti-concordant with the PFS in days. Results were ranked according to the C-index values calculated by comparing the predicted risk scores with the ground truth. This metric quantifies a model's ability in ranking the survival times based on the calculated individual risk scores, generalizing the Area Under the ROC Curve (AUC). It can account for censored data, i.e., when patients left the study after a given amount of time and encountered no event. The implementation is based on the Lifelines library[8] and adapted to handle missing values that are counted as non-concordant. It can be found on our GitHub repository[9] and was provided to the participants for transparency.

The final ranking across all participants was obtained by selecting for each team the best out of their possible 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 performed to avoid iterative overfitting.

4. Results

This section presents results in all three tasks in terms of challenge participation, algorithms descriptions, algorithms' performance, as well as task-specific analyses. The challenge was organized as a one-time event with fixed deadline (Sept. 14 2021). Only fully automatic methods were evaluated. The results and ranking are available on the AIcrowd platform. The Springer LNCS proceedings (Andrearczyk et al., 2022a) of the challenge include the overview paper (Andrearczyk et al., 2021b) and the 29 participants' papers. Each paper was reviewed by a minimum of two invited reviewers and one core organiser.

4.1. Task 1: Segmentation

For the segmentation task (Task 1), besides the standard participation and results, we report additional analyses to gain a better understanding of the algorithms' performance, their clinical value, and remaining challenges. In particular, we summarize and compare algorithms based on key design choices, we report results from ensembling the algorithms into a "super-algorithm", we evaluate simple automatic and semi-automatic PET thresholding methods as a performance baseline, the influence of tumor size and SUV on segmentation performance, the stability of the ranking and alternative rankings with bootstrap sampling, as well as the effect of reporting the best of 5 submissions per participant (e.g. risk of overfitting).

Participation. As of Sept. 14 2021 (submission deadline), we received a total of 181 submissions evaluated successfully for the first task. A total of 41 teams participated and 22 of them submitted a paper describing their methods and results. Only these teams that described their contribution in a paper submission were considered for the official ranking due to the limited scientific value of other results submissions.

Algorithms Summary. Almost all models use ensembles of 3D U-Nets. Attention modules were used by less than half of the participants without a clear benefit in terms of performance. Most methods pre-process the data with standard resampling, CT clipping and PET standardization. Various data-augmentations are used including rotation, scaling, flipping and noise addition. Some of the top-performing methods use SE normalization (Iantsen et al., 2020). The Dice loss was the most popular to train the models, combined with other losses such as cross-entropy and focal loss. Among the top five, only one approach did not use nnU-Net. Key elements of the top-performing algorithms are simple design choices including appropriate preprocessing, normalization, data augmentation and ensembling. The winning method (team "Pengy" Xie and Peng (2022)) used a well-tuned patch-based 3D

[8] `https://lifelines.readthedocs.io/en/latest/`, as of September 2022.
[9] `github.com/voreille/hecktor/tree/hecktor2020/src/evaluation`, as of September 2022.

nnU-Net with Squeeze and Excitation (SE) normalization Iantsen et al. (2021). A standard pre-processing and training scheme was used and the learning rate was adjusted dynamically using polyLR Chen et al. (2016). Five models were trained in a five-fold cross-validation with random data augmentation including rotation, scaling, mirroring, Gaussian noise and Gamma correction. The five test predictions are ensembled via probability averaging for the final results.

A detailed categorization of the methods in terms of pre-processing, data augmentation, model architecture, loss and training scheme can be found in Appendix B, Table B.7. More details on the individual methods can be found in the corresponding participants' papers.

Segmentation Performance. The results, including average DSC, HD95, precision, recall, F1-score and challenge rank are summarized in Table 2.

Team	DSC ↑	HD95 ↓	Precision ↑	Recall ↑	F1 ↑
Pengy (Xie and Peng, 2022)	**0.7785**	3.088	0.8361	0.7751	0.8044
SJTU EIEE 2-426Lab (An et al., 2022)	0.7733	**3.088**	0.8037	0.7972	0.8004
HiLab (Lu et al., 2022)	0.7735	3.088	0.7877	**0.8086**	0.7980
BCIOQurit (Yousefirizi and Rahmim, 2021)	0.7709	3.088	0.7974	0.8067	0.8020
Aarhus Oslo (Ren et al., 2022)	0.7790	3.155	0.8032	0.8085	**0.8058**
Fuller MDA (Naser et al., 2022a)	0.7702	3.143	0.8037	0.7925	0.7981
UMCG (De Biase et al., 2022)	0.7621	3.143	0.7881	0.7865	0.7873
Siat (Wang et al., 2022a)	0.7681	3.155	0.8211	0.7707	0.7951
Heck Uihak (Cho et al., 2022)	0.7656	3.155	0.7834	0.8036	0.7934
BMIT USYD (Meng et al., 2022)	0.7453	3.155	0.7980	0.7537	0.7752
DeepX (Yuan et al., 2022)	0.7602	3.270	0.7812	0.7996	0.7903
Emmanuelle Bourigault (Bourigault et al., 2022)	0.7595	3.270	**0.8746**	0.7133	0.7858
C235 (Liu et al., 2022)	0.7565	3.270	0.7774	0.7988	0.7880
Abdul Qayyum(Qayyum et al., 2022)	0.7487	3.270	0.7972	0.7586	0.7774
RedNeucon (Martinez-Larraz et al., 2022)	0.7400	3.270	0.7624	0.7877	0.7748
DMLang (Lang et al., 2022)	0.7046	4.026	0.8195	0.6647	0.7340
Xuefeng (Ghimire et al., 2022)	0.6851	4.193	0.7394	0.7199	0.7295
Qurit Tecvico (Salmanpour et al., 2022)	0.6771	5.421	0.6788	0.7318	0.7043
Vokyj (Juanco-Müller et al., 2022)	0.6331	6.127	0.7620	0.5935	0.6673
TECVICO Corp Family (Fatan et al., 2022)	0.6357	6.372	0.6355	0.7335	0.6810
BAMF health (Murugesan et al., 2022)	0.7795	3.057	0.8340	0.7706	0.8010
Wangjiao (Wang et al., 2022b)	0.7628	3.270	0.7977	0.7855	0.7916

Table 2: Summary of the segmentation results (Task 1). The official ranking is based on both DSC and HD95. Note that SJTU EIEE 2-426Lab is ranked second due to the HD95 slightly better than the third (HiLab), 3.0881603 vs 3.0881618. The two participants at the bottom were disqualified due to an excessive number of submissions on the 2020 data.

The results from the participants range from an average DSC of 0.633 to 0.779 and a median HD95 of 0.309 to 6.37. Note that SJTU EIEE 2-426Lab is ranked second due to the HD95 slightly better than the third (HiLab), 3.0881603 vs 3.0881618, and the ranking strategy described in 3.5. Statistical significance between pairs of teams was assessed with a one-tailed Wilcoxon test corrected for multiple hypotheses testing. For the DSC, no significant difference was found in the top-six group (i.e. Pengy vs SJTU, HiLab, BCIOQurit, Aarhus Oslo or Fuller MDA). The first statistically significant comparison was found when comparing Pengy with the 7th position UMCG (p = 0.0.022). For the HD95, the first significance was found when comparing Pengy with the 6th position Fuller MDA (p = 0.037). The precision and recall results, not used for the ranking, range from 0.6355 to 0.8746 and 0.5935 to 0.8086, respectively. All methods presented a similar trade off between precision and recall, with the exception of Emmanuelle Bourigault, which obtained the highest precision while being ranked only 12th due to a relatively low recall. The distributions of DSCs across patients and across participants are reported in Figures 2 and 3 respectively. The latter shows that some

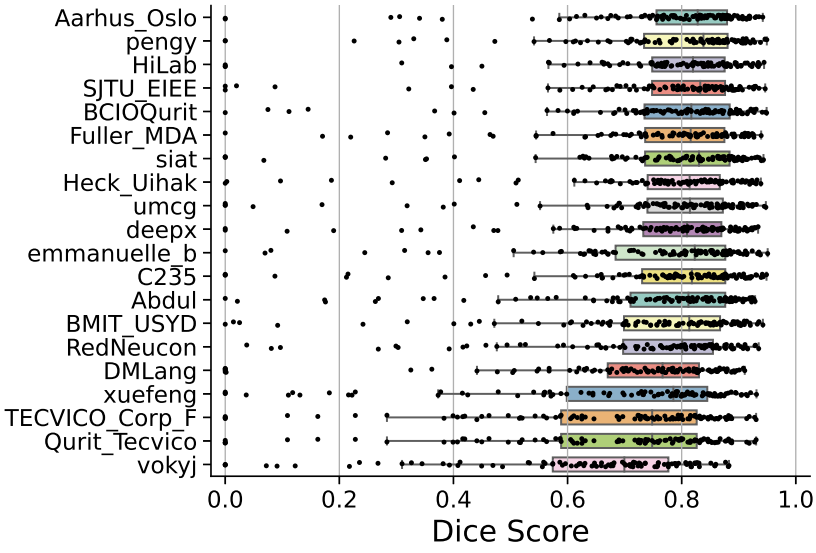


Fig. 2: Box plots of the distribution of the 101 test DSCs for each participant, ordered by decreasing DSC (different from the rank that combines DSC and HD95).

cases of both centers are incorrectly segmented by almost all algorithms (e.g. CHUV036, CHUV001, CHUP076). For CHUV036, for instance, a large metastatic lymph node is often incorrectly segmented as a primary tumor. For CHUV001, a metabolic volume at the level of the soft palate is incorrectly segmented by most algorithms, while the primary tumor situated lower than the average is commonly missed. Besides, some cases are correctly segmented by most algorithms (e.g. CHUV020 and CHUP029), while other cases, mostly in the CHUV center, show a large variation (e.g. CHUV016) across participants.

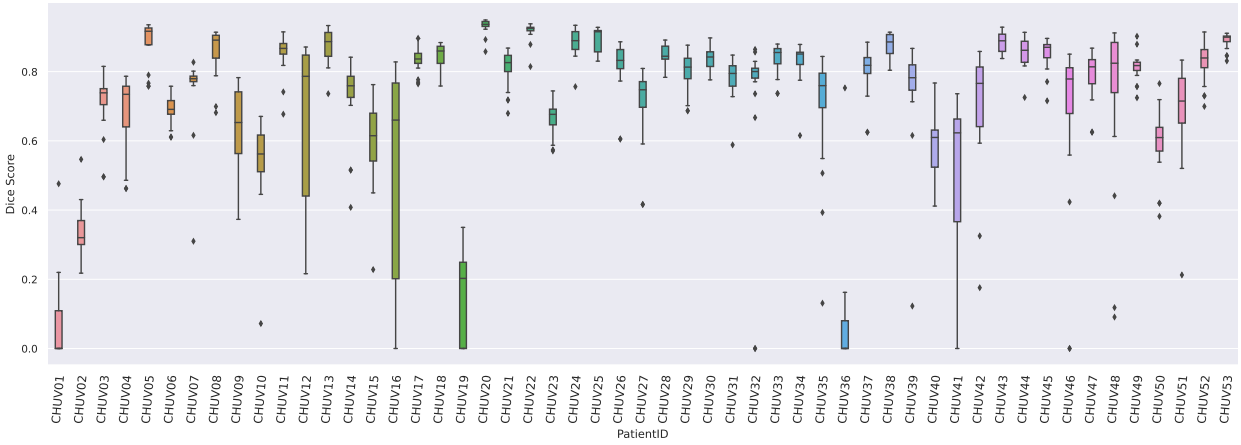
Ranking Robustness. The robustness of the ranking toward changes in the test set was assessed with a bootstrap analysis (n=1000). The methodology used here is inspired by the challengeR toolkit (Wiesenfarth et al., 2021). The results are reported in Fig. 8. We computed the Kendall rank coefficient between the actual rank and the ones obtained for each bootstrap. The following coefficients were obtained. Official rank (based on Borda count): 0.819 (0.744 - 0.885), average DSC: 0.843 (0.642 - 0.916), median HD95: 0.793 (0.689 - 0.882), and aggregated DSC: 0.841 (0.724 - 0.924). The aggregated DSC is defined as

$$\mathrm{DSC}_{\mathrm{agg}} = \frac{2\sum_i |A_i \cap B_i|}{\sum_i |A_i| + |B_i|}, \tag{3}$$

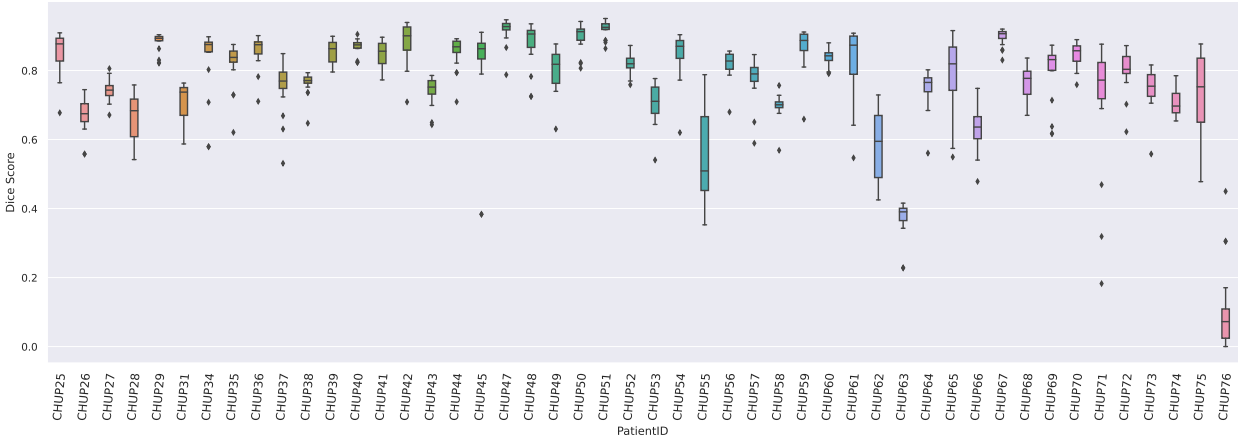
with A_i and B_i respectively the ground truth and predicted segmentation for image i, where i spans the entire test set. This metric was employed in Andrearczyk et al. (2022b) and is used as the ranking metric in HECKTOR 2022, for this reason we wanted to evaluate its stability. Overall, the ranking variability corroborates the statistical test which showed no significance among the top-performing teams.

Ensemble of Participants. In this section, we create a "super-algorithm" as an ensemble of the different participants' predictions. Such analyses often revealed superior performances to all submitted runs (Menze et al., 2014), leveraging the diversity of distinct methods (Hastie et al., 2009). We ensemble the (binary) predictions of multiple participants ((i) all 20 participants with paper submissions, and (ii) top-5 ranking participants) using the Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm (Warfield et al., 2004). A simpler ensembling method is also computed by taking the average of the different participants' predictions for each patient, and then thresholding at 0.5 to obtain a binary prediction. The results are reported in Table 3.

Most ensembles outperform the best participant result (pengy: DSC 0.778, HD95 3.09). The best ensemble performance is obtained by the average of all 20 participants with a DSC of 0.780 and HD95 3.06. Note that many participants already reported



(a) CHUV, mean DSC 0.714

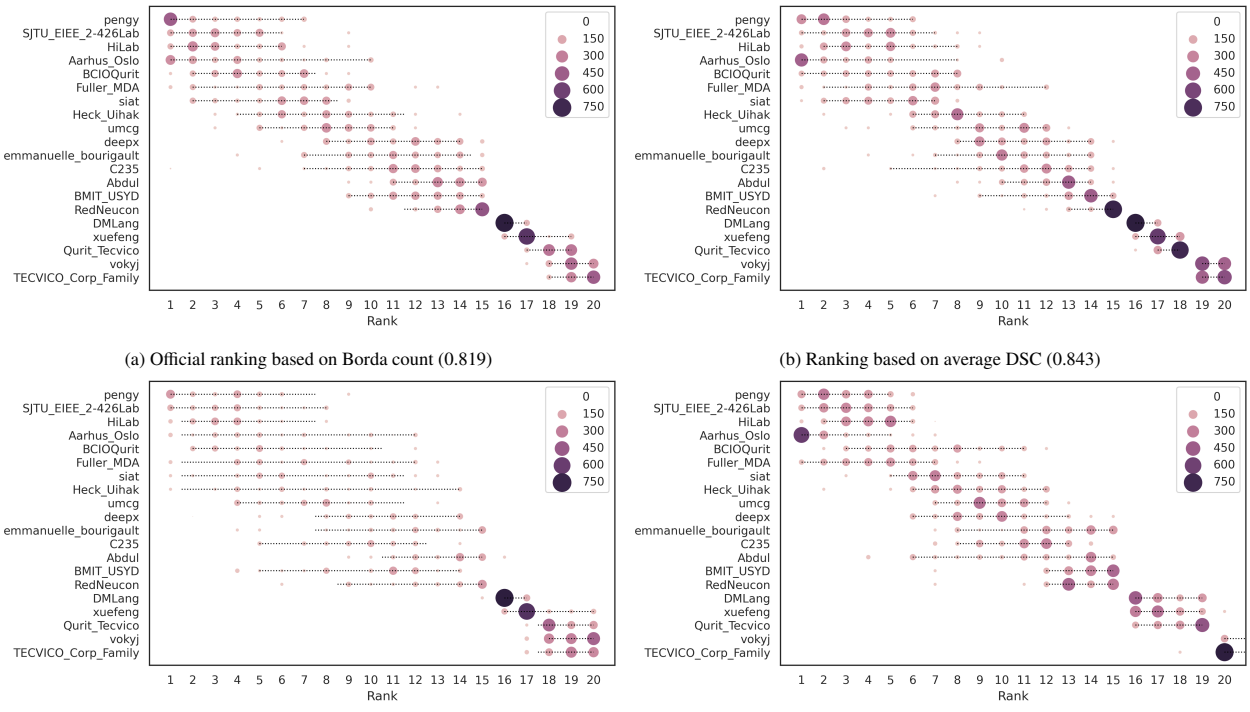


(b) CHUP, mean DSC: 0.765

Fig. 3: Box plots of the distribution of DSCs across the 20 participants for each of the 53 CHUV and 48 CHUP patients in the test set.

	top-5 STAPLE	top-20 STAPLE	top-5 average	top-20 average	pengy
DSC	0.779	0.758	0.779	**0.780**	0.778
HD95	**3.06**	3.27	**3.06**	**3.06**	3.0882

Table 3: Segmentation results of different ensembling methods and Team pengy (Xie and Peng, 2021), the winner of Task 1.



(a) Official ranking based on Borda count (0.819)

(b) Ranking based on average DSC (0.843)

(c) Ranking based on median HD95 (0.793)

(d) Ranking based on aggregated DSC (0.841)

Fig. 4: 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 Kendall tau is reported in brackets.

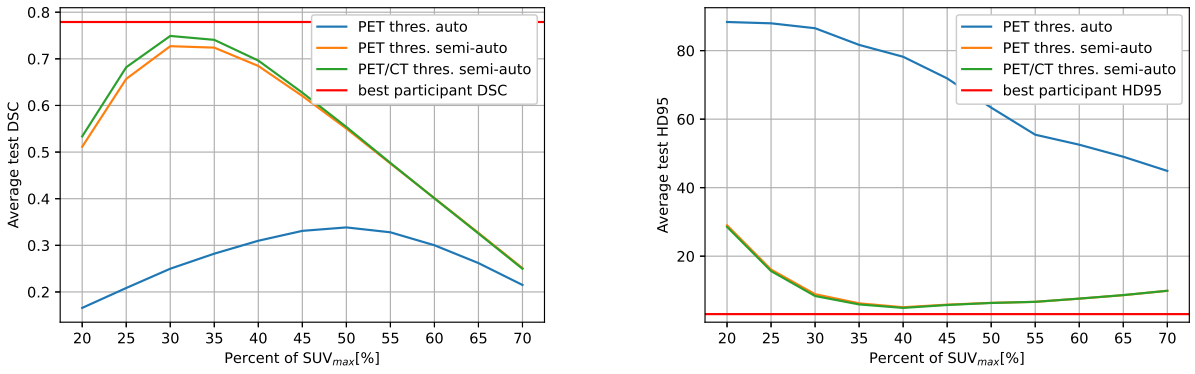
	All test		CHUV		CHUP	
	DSC	HD95	DSC	HD95	DSC	HD95
pengy (Xie and Peng, 2022)	0.778	3.09	0.766	3.27	**0.792**	**3.00**
STAPLE top-5	0.779	3.06	0.766	3.06	**0.794**	**3.00**
All participants (average)	0.738	3.30	0.714	3.52	**0.765**	**3.30**

Table 4: Comparison of performance across test centers (CHUV and CHUP).

results obtained by an ensemble of multiple independent network predictions (see Table B.7).

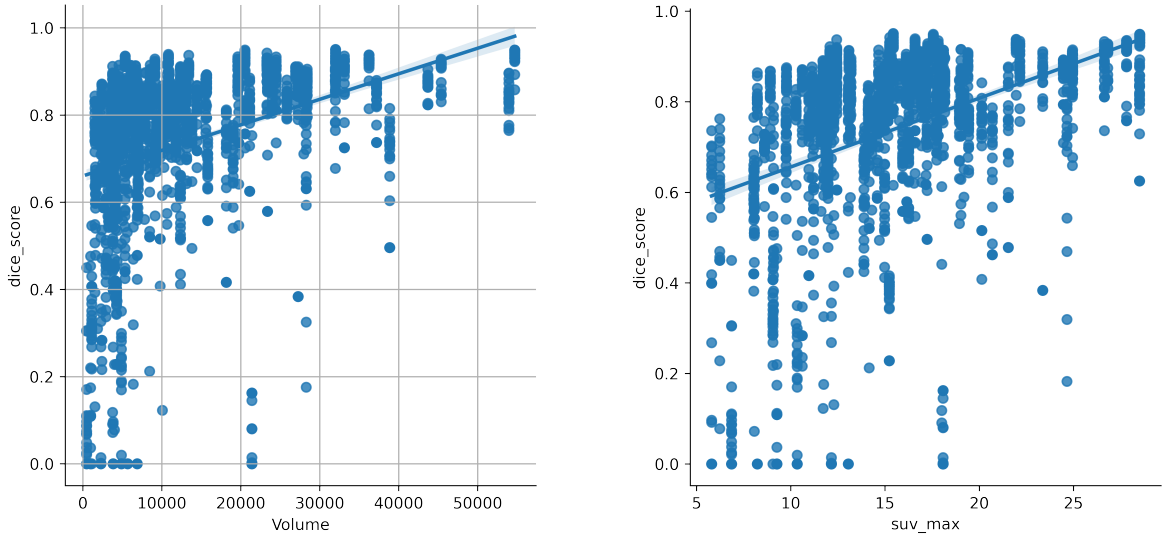
Inter-Center Performance. We compared the performances separately on the two centers subsets of the test set (CHUV and CHUP) for (i) the best participant (pengy), (ii) The STAPLE ensemble of the top-5 participants, and (iii) all 20 participants. The results are reported in Table 4. The results on the CHUV subset were lower than the CHUP one in the three scenarios, as discussed in Section 5.

PET Thresholding. PET thresholding is *de facto* the most widely used method for lesion segmentation in clinical routine, often computed after an an initial manual delineation of the field of interest. We compared the participants' results with simple automatic and semi-automatic PET thresholding methods. The fully-automatic threshold is obtained by thresholding the PET image at a given percentage of the maximum SUV value within the bounding-box (we evaluate a range of values). For the semi-automatic threshold, we mimic a manual indication of the tumor by an expert, followed by a similar threshold of the PET values. In practice, we first threshold the PET image, then extract the (26-)connected components and retain those (generally a single volume is retained) that overlap with the ground truth tumor volume. In Fig. 5, we report the results of these methods on the test set for various thresholds based on the percentage of the maximum SUV. Finally, we also report the results of the same semi-automatic thresholding with an additional threshold at -150 HU on the CT images to remove the air from the predictions.



(a) DSC (b) HD95

Fig. 5: Segmentation test performance (average DSC and median HD95) of PET thresholding-based method at different percentages of maximum SUV. The results of three methods are reported: the automatic PET threshold, the semi-automatic PET threshold (indicating the location of the ground truth GTVp), and the semi-automatic PET and CT (to remove the air) threshold. Best viewed in color.



(a) (b)

Fig. 6: Scatter plots of (a) DSC vs. tumor volume (voxel count in the VOI) and (b) DSC vs. SUVmax; for 20 participants. The corresponding Spearman rank correlations are 0.494, 0.428 respectively.

Influence of Tumor Size and SUV on 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 GTVp multiplied by the voxel volume. The Spearman rank correlation between the DSC and the tumor volume across all 20 participants and all tumors is 0.494 (p-value$<$ 0.001). In Fig. 6, we illustrate this correlation with a scatter plot of the DSCs as a function of the tumor size. We also evaluate the correlation of the DSC with the SUVmax (Spearman correlation 0.428). Fig. 7 relates the performance for each of the 20 algorithms for five ranges of tumor size. This figure was generated by grouping the 101 test cases into five bins (*i.e.* intervals) containing 21, 20, 20, 20, and 20 cases. The average DSC was then computed for each team in each bin.

Taking Best of Five Submissions: Risk of Overfitting ?. Each participant could upload up to 5 submissions on the test set during the challenge and the best result was used for the final ranking. The average number of submissions was 4.15 on the segmentation

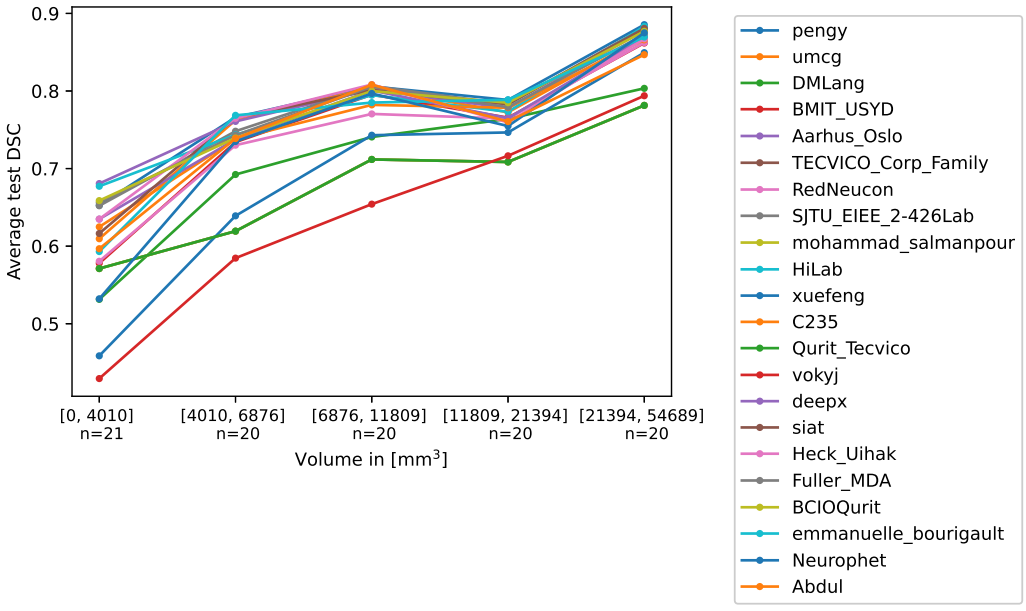


Fig. 7: Average DSC of each team's algorithm as a function of the volume of the tumors. This figure was generated by distributing the 101 test volumes in five bins of n =21, 20, 20, 20, and 20 each and then computing the average DSC for each bin. Best viewed in color.

task. Although this allows participants to evaluate different designs, it also rises the risk of overfitting the test data. We compared the average results of the participants when taking the best, the median, or the worst score of the participants' submissions. The average DSC across all best results of participants was 0.740, whereas it dropped to 0.721 and 0.681 when taking the median and minimum values respectively.

Please note that this analysis does not account for the false discovery rate inherent in scientific challenges, which may be influenced by the number of participants involved.

4.2. Task 2-3: Outcome Prediction

Participation. A total of 148 submissions were uploaded by 30 different teams for Task 2. However, out of 30 teams, only 17 were eligible for final ranking and prize by submitting a paper describing their algorithm. Task 3 was more complex to handle for participants as they had to encapsulate their algorithms in a Docker, which most probably explains the lower participation of seven teams (27 valid submissions, six papers). Of note, all these seven teams also participated in Task 2, which was of value because it allowed for a comparison between the use (or not) of reference contours.

Algorithms Summary. A detailed categorization of the methods in terms of pre-processing, segmentation step, image features, and model can be found in Appendix C, Table C.8. More details on the individual methods can be found in the corresponding participants' papers.

The methods proposed by the challengers can be categorized into three main approaches. The first one is to rely on the clinical variables only, thus deriving predictive models as a combination of clinical factors without exploiting the available PET/CT images. This was the case for two teams only. All the other teams exploited the PET/CT images to derive models. Amongst these, we could categorize two main methodological approaches. The first one is to rely on a "classical" radiomics approach, in which the tumor of interest is delineated, then engineered/handcrafted features are calculated to characterize this delineated volume, and finally a model is built by selecting some features to combine into a multiparametric model through one or several algorithms. Note that a simplified case of such an approach is to calculate and use only one or a couple of features, e.g., the tumor volume (and/or shape),

which three teams chose to do. The last main approach relies on DL models, using either pre-trained convolutional networks to extract "deep features" that are subsequently used for modeling through learning algorithms, or to train a deep network specifically in an end-to-end manner.

In both the standard radiomics and the DL approaches, PET/CT images can be exploited in different ways. For instance, four teams chose to implement a fusion of the two modalities into a single resulting image to use as an input, whereas other teams chose to keep the PET and the CT image as separate inputs of their pipelines. Only one team chose to exploit both the fusion and separate modalities. Some challengers also chose to include the segmentation mask as an additional input. Only three teams did not rely on segmentation of the tumor at all, and most of them (12 teams) relied on the output of their participation in Task 1 to get a tumor segmentation mask, which also allowed some of them participating in Task 3 to carry out a comparison between models obtained by relying on the expert reference contours or an automatic segmentation.

Both approaches also require different strategies to include clinical variables. All challengers that developed models using radiomics features or DL architectures chose to include, in one way or another, the available clinical factors in their models. Some even implemented imputation mechanisms in order to fill in missing values (e.g., the HPV status). Only four teams relied on the ensembling of several models, three of them being in the top-ranked teams. This ensembling was carried out by relying on a consensus or average of the networks' outputs in the 10 folds of cross-validation Naser et al. (2022b), a voting of eight different models (trained through ML algorithms exploiting handcrafted features) Salmanpour et al. (2022), ensembling 10 networks (five trained using a leave-one-center- out cross-validation and five trained using a 5-fold cross-validation) Meng et al. (2022) or averaging 18 trained networks, the best three of each of six folds of cross-validation Ma and Yang (2021).

The winner of the challenge (team "BiomedIA") in Task 2 did not participate in Task 3. It first carried out a comparison of the prediction performance achievable by relying on either all clinical variables with imputing missing values, or only on the ones with values available for all patients. They determined that better prediction was achieved using only variables with complete values. Then, they generated a new fused PET/CT image for the input of their pipeline by averaging the PET and CT modalities. This new PET/CT fusion was further cropped in order to focus on the tumor area, testing two different sizes (50x50x50 and 80x80x80). Better results were obtained with the larger area. They then trained a 3D CNN (Deep-CR) to extract "deep" features from either the fused PET/CT (one path) or the PET, the CT and the fused PET/CT (three separate paths). The authors exploited the OPTUNA (Akiba et al., 2019) framework to determine the best hyperparameters such as the kernel sizes and the number of layers. The resulting network consisted of two blocks, each block containing two 3D convolutional, ReLU activation and batch normalization layers. These 3D CNN blocks (kernel sizes 3 and 5, 32, 64, 128 and 256 output channels) are followed by 3D max pooling layers. The two feedforward layers contain 256 neurons each. The batch size, learning rate, and dropout were experimentally set to 16, 0.016, and 0.2 respectively for the training, for 100 epochs using the Adam optimizer. The obtained "deep" features, along with the clinical variables, were then fed into a Multi-Task Logistic Regression (MTLR) algorithm. MTLR consists of a sequence of logistic regression models created at various time points in order to evaluate the probability of an event. The authors integrated neural networks into the MTLR process in order to achieve non-linearity. No cross-validation or data augmentation was used. Of note, the results of 3D CNN and MTLR (i.e., exploiting both images and clinical variables) were averaged with the prediction of a Cox model using only clinical variables to obtain the best result.

The team "Fuller MDA", ranked second in Task 2 and first in Task 3 also implemented a pipeline approach relying on DL. They elected to choose only clinical variables without missing values and to encode them into an image matrix, allowing to feed it along with the PET and CT images (original bounding-boxes without further cropping) as separate channels to a DenseNet121

Team	C-index Task 2	C-index Task 3
BioMedIA Saeed et al. (2022)	**0.7196**	na
Fuller MDA Naser et al. (2022b)	0.6938	**0.6978**
Qurit Tecvico Salmanpour et al. (2022)	0.6828	na
BMIT_USYD Meng et al. (2022)	0.6710	na
DMLang Lang et al. (2022)	0.6681	na
TECVICO_C. Fatan et al. (2022)	0.6608	na
BAMF Health Murugesan et al. (2022)	0.6602	0.6602
ia-h-ai Starke et al. (2022)	0.6592	0.6592
Neurophet Lee et al. (2022)	0.6495	na
UMCG Ma et al. (2022)	0.6445	0.6373
Aarhus Oslo Huynh et al. (2022)	0.6391	na
RedNeucon Martinez-Larraz et al. (2022)	0.6280	na
Emmanuelle B. Bourigault et al. (2022)	0.6223	na
BCIOQurit Yousefirizi et al. (2022)	0.6116	0.4903
Vokyj Juanco-Müller et al. (2022)	0.5937	na
Xuefeng Ghimire et al. (2022)	0.5510	0.5089
DeepX Yuan et al. (2022)	0.5290	na

Table 5: Summary of the outcome prediction results. “na” stands for “not available. All participants of task 3 also participated in task 2.

CNN. This CNN contained 6, 12, 24, and 16 repetitions of dense blocks, each dense block containing a pre-activation batch normalization, ReLU, and a $3 \times 3 \times 3$ convolution followed by a batch-normalization, ReLU, and $1\times1\times1$ convolution. The model has two (PET, CT) or three (+clinical) input channels, each of size $144 \times 144 \times 144$ and 20 output channels, as the PFS was discretized into 20 discrete intervals. It was trained through a 10-fold cross-validation scheme (non-overlapping folds), with data augmentation (random horizontal flips of 50%, random affine transformations with an axial rotation range of 12° and a scale range of 10%), for 800 iterations with a decreasing learning rate, the Adam optimizer and a negative log-likelihood loss. The models obtained through the 10 folds were then ensembled with two different approaches, consensus or averaging. For consensus, the mean conditional probability survival vector was first estimated by getting the mean value for each time interval, and then by computing the cumulative survival probability for each interval to estimate the consensus PFS values from the 10 models. For averaging, the PFS was estimated for each patient by each model and the mean value of the 10 predicted PFS values was calculated.

Performance. The results of Tasks 2 and 3 are reported in Table 5. First of all, it is important to emphasize that the best performance obtained among the challengers using only the clinical variables was 0.649 (”Neurophet”). In comparison, we also performed baseline results obtained with the tumor volumes and SUVmax, resulting in a C-index of 0.5683 and 0.5722, respectively. These results are significantly lower than the top-performing teams results. Looking at the performances obtained by the challengers, eight of them produced C-index values close to or lower than this (between 0.529 and 0.644). Among these, four did not rely (either fully or at all) on the PET/CT images. The team ”Aarhus Oslo” used only the clinical variables (C-index of 0.639). The three other teams exploited the images only to calculate a tumor volume, combined (”Vokyj”, 0.594 and ”RedNeucon”, 0.628) or not (”Xuefeng”, 0.551) with clinical variables.

Of note, the team ”BioMedIA” (who reached the first rank by combining clinical variables with imaging information) also evaluated a baseline model using only the clinical variables, which reached a higher C-index of 0.66. All eight teams achieving a higher performance exploited the quantitative content of both PET and CT images as input (either relying on some kind of fusion of both modalities or exploiting each image separately), and also included clinical variables, in one way or another. Amongst these eight teams, the ones relying on DL approaches occupied four of the five first places of the final ranking. The ones relying on a more

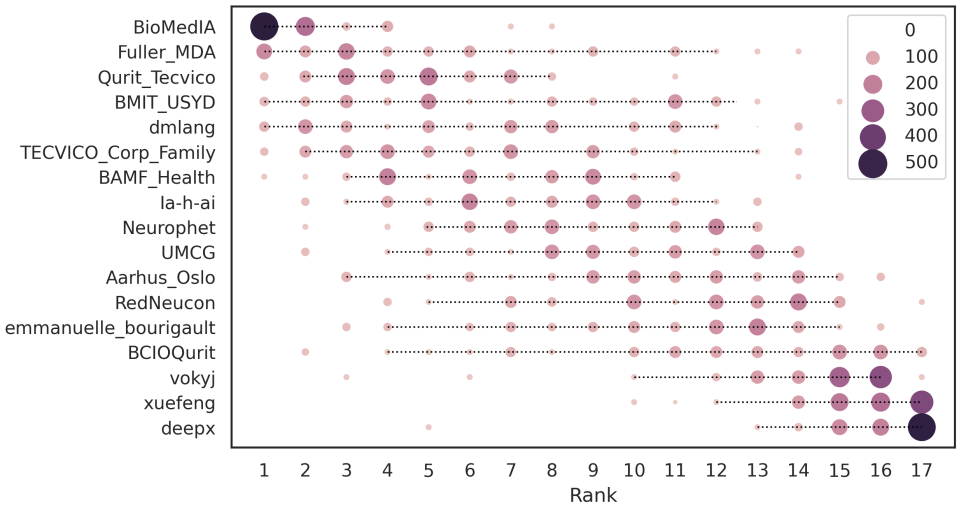
classical radiomics approach (i.e., extraction of handcrafted features from a delineated tumor volume, subsequently selected and combined in a model through ML algorithms) were ranked 3rd and 6th-8th. Looking at the final ranking, it can be deduced that the best results were obtained by pipelines relying on a DL approach without using segmentation of the tumor (i.e., the initial input in the pipeline is a bounding-box containing the tumor and its surroundings), combined with clinical variables. The best performance in Task 2 was achieved by the team "BiomedIA". Their proposed framework relying on one path network (using only the fused PET/CT image as input) led to the best performance of 0.720 whereas the version using 3 paths (PET, CT and PET/CT) achieved a lower C-index of 0.67. Their baseline MTLR framework relying on clinical variables only achieved a C-index of 0.66. As their pipeline did not rely at all on provided segmentation of the tumor, it was logical not to participate in Task 3. Statistical significance between pairs of teams was assessed using the method of Kang et al. (2015). No significant difference was found in the top-five group (i.e. BioMedIA vs FullerMDA, Qurit Tecvico, BMIT USYD or DMLang). The first statistically significant comparison was found when comparing BioMedIA with the 6th position TECVICO_C ($p = 0.004$).

The best performance in Task 3 was obtained by the team ranked second in Task 2, "FullerMDA". Their best result in Task 2 (0.694, rank 2nd) was obtained with a model exploiting both image and clinical, with a consensus approach to aggregate the 10 models of the cross-validation scheme. Using the average approach instead lowered the C-index to 0.689, whereas not using the clinical information lowered the performance even lower, with C-index of 0.645 and 0.651 with average and consensus respectively. In Task 3, they used ground truth masks as an additional input channel to the same network, achieving a C-index of 0.696 and 0.698 (average and consensus respectively). Using the test of Kang et al. (2015), no significant difference was found in the top three group (i.e. FullerMDA vs BAMF Health or ia-h-ai). The first statistically significant comparison was found when comparing FullerMDA with UMCG ($p = 0.001$).

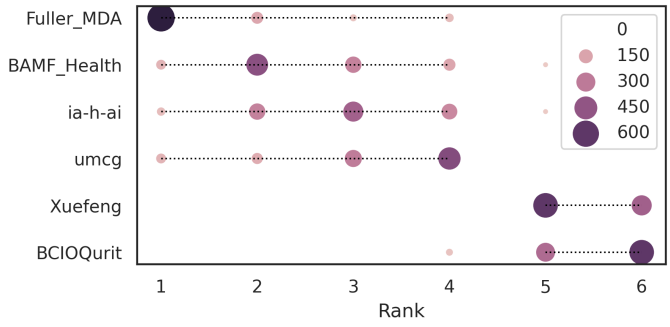
Inter-Center Performance. We evaluated the performance of the best team (BioMedIA, with a C-index on the entire test set of 0.720) separately on the two centers subsets of the test set. The C-index on CHUV was much lower (0.648) compared to the value of 0.727 obtained on the CHUP test. This better prediction for CHUP patients can be explained by the fact that some CHUP patients were included in the training set, whereas all CHUV patients were held out for the test set. This was observed even though the survival profiles of CHUP patients were significantly different (a higher rate of events and shorter PFS overall) compared to all five other centers (thus, also compared to the CHUV set in the testing data). The lower generalization of performance on the CHUV dataset is therefore more likely explained by differences in PET/CT image properties than on survival profiles.

Ensembling. In order to evaluate the potential complementary power of the different best predictions, we calculated a mean and median ranking of patients based on the predicted scores of the five best results ranging from 0.718 to 0.668 (BioMedIA, FullerMDA, QuritTecVico, BMITUSYD, and DMLang). We chose these 5 results as they are those with a C-index value at least above 0.665, compared to the prediction performance obtained using a simple model relying on clinical variables only (C-index 0.64-0.65). Of note, all these five results were obtained through a DL based pipeline, except QuritTecVico. Such an ensembling ranking only marginally improved the final prediction, with C-index values of 0.724 and 0.728 for the mean and median of ranks respectively.

Ranking Robustness. Looking at the robustness of the results (Figure 8), even though it overall confirms the hierarchy between the challengers, we observed quite a large variability in the ranking amongst the teams according to the bootstraps of the test data, as assessed by an overall Kendall rank coefficient for Task 2 of 0.742 (0.618 - 0.868) and for Task 3 of 0.739 (0.333 - 1.00). For instance, the team "BioMedIA" that won Task 2 was ranked 1st only in half of the 1000 bootstraps, its ranking going as low as 8th for a few of the bootstrapped instances (between rank 1 and 5 for the 95% confidence interval). Another example: the team



(a) Ranking for Task 2



(b) Ranking for Task 3

Fig. 8: 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.

"Neurophet", ranked 8th in 200 out of 1000 bootstrapped instances, was ranked between 2nd and 12th ranks (95% confidence interval) with similar numbers of instances (≈50-200). These results corroborate the statistical tests reported for tasks 2 and 3, without significance between the top performing teams.

Taking Best of Five Submissions: Risk of Overfitting ?. Each participant could upload up to five submissions on the test set during the challenge. The average number of submissions was 4.82; 16 participants uploaded five submissions and one uploaded only two. We thus compared the average results of the participants when taking the best, the median, or the minimum score of the participants' submissions. The average C-index across all best results of participants was 0.640, whereas it decreased to 0.598 and 0.523 when taking the median and minimum values respectively. These results are less relevant than for Task 1 (Section 4.1) because some participants submitted concordant and anti-concordant results of the same predictions (e.g. BCIOQurit, deepx). This is due to the fact that these participants first submitted concordant results, e.g. time to recurrence, then realized that anti-concordant results were required and re-submitted inverted scores. The C-indexes for these results are thus $C_{concordant} = 1 - C_{anti-concordant}$, which makes the above analysis biased with small non-relevant values. We did not remove C-indexes below 0.5 for this analysis because one cannot know for sure when the participants' predictions are voluntarily concordant or not.

One may also wonder whether the results are better than random predictions due to this best algorithm selection with many submissions. To answer this question, we approximated the sampling of the best C-index under the null hypothesis H0: all $16 \times 5 +$

2 = 82 submissions are independent and identically distributed random predictions. We generated random predictions in the range [0, 1] for all 82 submissions, and took the best C-index as the "winner" of these random predictions. We repeated this process 1000 times to report the distribution of these 1000 random winners' C-indices. The average was 0.617 (vs. 0.720 of the actual winner) ± 0.021 standard deviation. The probability of a random winner obtaining a C-index above 0.720 (1st rank) was < 0.001. Note that the independence of the submissions assumed in the null hypothesis overestimates the probability of the latter, i.e., the more random submissions, the more likely the best random C-index will be high. In practice, we observed various teams submitting concordant and anti-concordant scores.

5. Discussions

Task 1 - Segmentation. Task 1 on GTVp segmentation was conducted for the second time after HECKTOR 2020. This year, data from a new center (CHUP) was added. Some of the trends observed in the first edition were confirmed with this larger dataset and increased participation. The winner method (Xie and Peng, 2022) was similar to the previous year's winning approach (Iantsen et al., 2021). Following the general trend in medical image segmentation, DL methods based on U-Net models were mostly used in the challenge, as reported in Table B.7. Model ensembling, as well as data preprocessing and augmentation, seem to have played an important role in achieving top-ranking results. Note that, as reported in Table 3, ensembles of participants' results outperformed the top-1 performance. Four of the top-5 teams used the nnU-Net framework (Isensee et al., 2021), reflecting the high performance obtained by this method in various segmentation challenges and tasks. We evaluated a vanilla nnU-Net trained on the training set. The only preprocessing applied to the data was to crop the images to the provided bounding-box. This model achieved a DSC of 0.770 and a HD95 of 3.27 mm and would have ranked 10$^{\text{th}}$, proving a strong starting point. The modifications made by the participants were relevant and further improved its performance. Note that the vanilla nnU-Net was mainly hindered by its poor HD95, a metric which was not used in the elaboration of the nnU-Net framework.

The ranking relied on the Borda counting based on average DSC and median (because not bounded) HD95. The results of the HD95 are mainly influenced by the voxel spacing in the inferior/superior axis since it is larger than the voxel spacing in the axial plane. The robustness of the HD95 ranking is therefore lower than the average DSC, as shown in Figure 4. The final ranking is, however, stable with a Kendall tau coefficient > 0.8. The aggregated DSC is also a stable metric and will be considered for HECKTOR 2022 since it is appropriate for the future task of lymph node segmentation, with cases without target volumes.

The inter-center performance difference was reported in Table 4, with results in the CHUV cases lower than those in the CHUP cases. Two main reasons may explain this large difference: (i) CHUP cases are present in the training set and (ii) CHUP tumors are larger (and slightly more metabolic), on average, which biases the DSC to higher values. The median volumes of the CHUV and CHUP test cases are 7017 and 10879 mm^3, respectively. The means are 13305±13468 and 13350±10025 mm^3. The mean SUVmax in the GTVp are 15.2 and 15.6, respectively. The correlation of performance with the tumor size and SUVmax is reported in Figures 6 and 7.

Automatic PET thresholding, commonly used in clinical routine, was evaluated in Figure 5. The best results, with an average DSC of 0.749 and HD95 8.37, are obtained with a 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). These non-fully automatic results remain lower than most algorithms' results.

Since the CHUV cohort was used both in HECKTOR 2020 and 2021, a risk of cheating existed even though we never disclosed the ground truth contours for these cases. A first risk lied in the possibility for 2020 participants to manually annotate the CHUV

cases, which could not be assessed. A second possibility to cheat was to submit several runs on the leaderboard of HECKTOR 2020. This was carefully monitored, and two teams were subsequently disqualified (see bottom rows of Table 2).

Finally, we observed a minor overestimation of the participants' performance due to the process of reporting the best out of five possible submissions. This selection resulted in a marginal performance increase over a median selection (DSC 0.740 vs 0.721).

Tasks 2 and 3: Outcome prediction. Tasks 2 and 3 on outcome prediction were first put in place for this 2021 HECKTOR edition. Current standard of care for estimating outcome is often based on TNM staging. However, the patients included in HECKTOR are quite homogeneous in terms of stage as they were all treated with concomitant radiochemotherapy. Therefore it is particularly relevant to evaluate the prognostic power of image-based AI models in order to identify outcome beyond TNM, knowing that several image- and basic clinical- based prognostic biomarkers were reported in the literature (Castelli et al., 2019; dit Deprez et al., 2022; Morand et al., 2018). It is worth noting that a negative finding for SUVmax was also reported (Patel et al., 2021).

Tasks 2 and 3 attracted a promising number of submissions, with varied pipeline and algorithm developments. One of the most interesting findings was the comparison between the three main categories of approaches, namely a first relying on clinical factors only, and two relying on the available PET/CT images, with the use of either DL or more classical radiomics modeling approaches. Despite the relatively limited training dataset size (224 patients) compared to other computer vision tasks where DL has established itself as the state of the art thanks to the availability of thousands or even millions of images for training, the challengers relying on DL methods tended to slightly outperform the ones relying on radiomics approaches using handcrafted features selected and combined through classical ML techniques. Indeed, DL approaches took four of the five first places in the ranking.

Another important finding is that, despite their best efforts, the challengers failed to improve the predictive performance by a large margin thanks to the use of the PET/CT images (C-indices ranging from 0.65 to 0.72), compared to a basic model relying on clinical factors alone (C-index of ≈0.64-66). Half of the challengers even achieved lower performance, hinting at a lack of generalizability of their trained model when applied to the test set. It remains to be investigated whether further improving the C-index is possible with more appropriate or novel methodological developments, and / or if this can be achieved only by making a larger training dataset available, which will be the case in the 2022 HECKTOR edition (489 patients for training instead of 224).

A third important observation is that the dataset we used in HECKTOR 2021 offered an important challenge, related to its multicentric nature. Indeed, PET/CT images came from six different centers (five in the training set, two in the test set, one of which was also present in the training set) with five different scanner models from the three main vendors (Philips, GE, Siemens), associated with various acquisition settings and reconstruction parameters. However, none of the teams implemented explicit harmonization techniques (of the images or the features) beyond classical algorithms such as image interpolation, rescaling and normalization. Further harmonization (of images and/or features) might have helped get more generalizable models, especially in the case of standard radiomics modeling relying on handcrafted features, for which the impact of imaging characteristics is now well documented. This could also contribute to explaining the relative higher performance obtained by challengers using DL pipelines where pre-processing steps (interpolation, normalization, etc.) are commonly implemented and have a side-effect of filtering out some differences in images characteristics. By comparison, approaches relying on handcrafted radiomic features that are notoriously sensitive to the multicentric nature of the images might be at a disadvantage without explicit harmonization procedures. Such statistical harmonization, for example using the ComBat approach (Johnson et al., 2007), could help in improving the performance of models based on handcrafted radiomic features as shown in (Abdallah et al., 2022). This aspect will become even more important in the 2022 edition of HECKTOR, in which 489 cases from seven centers are available for training, and 339 cases from two additional centers will constitute the testing set.

Last but not least, it was observed by several challengers that using alternative (mostly larger) volumes of interest led to better models compared to exploiting reference ground truth contours. This observation is in line with previously reported studies that used peri-tumoral regions for outcome prediction (Leger et al., 2020). This can be seen in the overall performance of challengers in Task 2 where they had to rely on either their automatic segmentation results from Task 1 or alternative inputs to the characterization pipeline (including, for the best results, the entire bounding-box without segmentation), compared to their performance in Task 3 where they had ground truth contours available, and where they often reported better results using other segmentation or inputs.

In terms of challenge design and ranking approach, a few important points can be noted and were taken into account for the organization of the next installment in 2022.

First, a minor over estimation of the participants' performance is associated with the process of reporting the best out of five possible submissions. Although this does not invalidate the results and we specifically asked the challengers to report the method of their best result in their paper, we might consider other strategies, such as reporting the mean of the submissions in the future.

Second, the inter-center performance variation with lower performance observed on the data not represented in the training set compared to the ones of the shared center further emphasizes the potential lack of generalizability of the developed models. It thus seems such predictive models still need far more work before being largely used in a real clinical setting. In the HECKTOR 2022 installment, all data in the testing set will be from several centers not present in the training set.

Third, the variability of rankings amongst the teams according to bootstrapping of the test data emphasizes on the relatively small differences of performance between the submitted pipelines, which is also highlighted by the lack of statistical significance among the top five teams for Task 2.

6. Conclusion

This paper presented the HECKTOR 2021 challenge data, participation and ranking as well as an extensive analysis of the results. A strong participation was observed in the tasks of primary tumor segmentation and patient outcome prediction from PET/CT images and clinical data.

In the first task, the segmentation results were marginally improved compared to the first challenge edition (Oreiller et al., 2022). Simple designs based on nnU-Net obtained the best results (above the baseline nnU-Net). The participants' algorithms largely outperformed simple PET thresholding methods. The performance was strongly influenced by the tumor size and, to a lower extent, by the SUVmax, i.e. higher performance related to larger and more metabolic tumors.

In the outcome prediction tasks, DL algorithms obtained the best results, leveraging information from the images and clinical data. With a C-index >0.7 the algorithms showed potential capabilities to model the progression of tumors in internal and external data. Surprisingly, the best results were also obtained without the use of ground truth tumor delineations, opening the door to large-scale studies without the need for costly manual annotations.

As future work, the third edition of HECKTOR at MICCAI 2022 will feature a larger dataset with additional centers. In order to reach fully-automatic pipelines, bounding-boxes locating the oropharyngeal regions will no longer be provided. Besides, the segmentation task will combine primary tumor and metastatic lymph nodes segmentation due to their prognostic value for the final outcome prediction.

Acknowledgments

We thank all participants for their contributions in the HECKTOR 2021 challenge. We thank Siemens Healthineers[10] Switzerland, Aquilab[11] and Bioemtech[12] for kindly sponsoring Tasks 1, 2 and 3, respectively, with a prize of 500€each. We thank MIM software[13] for kindly providing licences to coordinate VOI contouring in PET/CT images. 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), the Hasler Foundation and the Swiss Cancer Research foundation with the project TARGET (KFS-5549-02-2022-R).

References

Abdallah, N., Xu, H., Marion, J.M., Tauber, C., Carlier, T., Chauvet, P., Lu, L., Hatt, M., 2022. Predicting progression-free survival from FDG PET/CT images in head and neck cancer : comparison of different pipelines and harmonization strategies in the HECKTOR 2021 challenge dataset, in: Proceedings of the IEEE NSS-MIC.

Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 5, 1–9.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

An, C., Chen, H., Wang, L., 2022. A coarse-to-fine framework for head and neck tumor segmentation in CT and PET images, in: Lecture Notes in Computer Science (LNCS) Challenges.

Andrearczyk, V., Fontaine, P., Oreiller, V., Castelli, J., Jreige, M., Prior, J.O., Depeursinge, A., 2021a. Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer, in: International Workshop on PRedictive Intelligence In MEdicine, Springer. pp. 147–156.

Andrearczyk, V., Oreiller, V., Boughdad, S., Rest, C.C.L., Elhalawani, H., Jreige, M., Prior, J.O., Vallières, M., Visvikis, D., Hatt, M., et al., 2021b. Overview of the hecktor challenge at miccai 2021: automatic head and neck tumor segmentation and outcome prediction in pet/ct images, in: 3D Head and Neck Tumor Segmentation in PET/CT Challenge, Springer. pp. 1–37.

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

Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A., 2022a. Head and neck tumor segmentation and outcome prediction .

Andrearczyk, V., Oreiller, V., Jreige, M., Castelli, J., Prior, J.O., Depeursinge, A., 2022b. Segmentation and classification of head and neck nodal metastases and primary tumors in pet/ct, in: 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).

Andrearczyk, V., Oreiller, V., Jreige, M., Vallières, M., Castelli, J., Elhalawani, H., Boughdad, S., Prior, J.O., Depeursinge, A., 2020b. Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT, in: 3D Head and Neck Tumor Segmentation in PET/CT Challenge, Springer. pp. 1–21.

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

Bourigault, E., McGowan, D.R., Mehranian, A., Papiez, B.W., 2022. Multimodal PET/CT tumour segmentation and prediction of progression-free survival using a full-scale UNet with attention, in: Lecture Notes in Computer Science (LNCS) Challenges.

Burke, H.B., Goodman, P.H., Rosen, D.B., Henson, D.E., Weinstein, J.N., Harrell Jr, F.E., Marks, J.R., Winchester, D.P., Bostwick, D.G., 1997. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 79, 857–862.

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? Clinical nuclear medicine 42, e281–e285.

Castelli, J., Depeursinge, A., Devillers, A., Campillo-Gimenez, B., Dicente, Y., Prior, J., Chajon, E., Jegoux, F., Sire, C., Acosta, O., et al., 2019. PET-based prognostic survival model after radiotherapy for head and neck cancer. European journal of nuclear medicine and molecular imaging 46, 638–649.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. CoRR abs/1606.00915.

Cho, M., Choi, Y., Hwang, D., Yie, S.Y., Kim, H., Lee, J.S., 2022. Multimodal spatial attention network for automatic head and neck tumor segmentation in FDG-PET and CT images, in: Lecture Notes in Computer Science (LNCS) Challenges.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34, 187–202.

Dang, M., Lysack, J., Wu, T., Matthews, T., Chandarana, S., Brockton, N., Bose, P., Bansal, G., Cheng, H., Mitchell, J., et al., 2015. Mri texture analysis predicts p53 status in head and neck squamous cell carcinoma. American Journal of Neuroradiology 36, 166–170.

De Biase, A., Tang, W., Sourlos, N., Ma, B., Guo, J., Sijtsema, N.M., van Ooijen, P., 2022. Skip-SCSE multi-scale attention and co-learning method for oropharyngeal tumor segmentation on multi-modal PET-CT images, in: Lecture Notes in Computer Science (LNCS) Challenges.

dit Deprez, L.W.L., Morand, G.B., Thüring, C., Pazahr, S., Hüllner, M.W., Broglie, M.A., 2022. Suvmax for predicting regional control in oropharyngeal cancer. European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery 279, 3167–3177.

Diamant, A., Chatterjee, A., Vallières, M., Shenouda, G., Seuntjens, J., 2019. Deep learning in head & neck cancer outcome prediction. Scientific reports 9, 1–10.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

[10] `https://www.siemens-healthineers.com`, as of September 2022.

[11] `https://www.aquilab.com/`, as of September 2022.

[12] `https://bioemtech.com/`, as of September 2022.

[13] `https://www.mimsoftware.com/`, as of September 2022.

Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Godau, P., Cheplygina, V., Kozubek, M., Ali, S., et al., 2022. Biomedical image analysis competitions: The state of current participation practice. arXiv preprint arXiv:2212.08568 .

Fatan, M., Hosseinzadeh, M., Askari, D., Sheykhi, H., Rezaeijo, S.M., Salmanpoor, M.R., 2022. Fusion-based head and neck tumor segmentation and survival prediction using robust deep learning techniques and advanced hybrid machine learning systems, in: Lecture Notes in Computer Science (LNCS) Challenges.

Fontaine, P., Andrearczyk, V., Oreiller, V., Castelli, J., Jreige, M., Prior, J.O., Depeursinge, A., 2021. Fully automatic head and neck cancer prognosis prediction in PET/CT, in: International Workshop on Multimodal Learning for Clinical Decision Support, Springer. pp. 59–68.

Foster, B., Bagci, U., Mansoor, A., Xu, Z., Mollura, D.J., 2014. A review on segmentation of positron emission tomography images. Computers in biology and medicine 50, 76–96.

Ghimire, K., Chen, Q., Feng, X., 2022. Head and neck tumor segmentation with deeply-supervised 3D UNet and progression-free survival prediction with linear model, in: Lecture Notes in Computer Science (LNCS) Challenges.

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

Harrison, K., Pullen, H., Welsh, C., Oktay, O., Alvarez-Valle, J., Jena, R., 2022. Machine Learning for Auto-Segmentation in Radiotherapy Planning. Clinical Oncology 34, 74–88. doi:10.1016/J.CLON.2021.12.003.

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

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584.

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. Medical Image Analysis 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 transactions on medical imaging 28, 881–893.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

Huynh, B.N., Ren, J., Groendahl, A.R., Tomic, O., Korreman, S.S., Futsaether, C.M., 2022. Comparing deep learning and conventional machine learning for outcome prediction of head and neck cancer in PET/CT, in: Lecture Notes in Computer Science (LNCS) Challenges.

Iantsen, A., Visvikis, D., Hatt, M., 2020. Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined pet and ct images, in: 3D Head and Neck Tumor Segmentation in PET/CT Challenge, Springer. pp. 37–43.

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, in: Lecture Notes in Computer Science (LNCS) Challenges.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8, 118–127.

Juanco-Müller, Á.V., Mota, J.F.C., Goatman, K., Hoogendoorn, C., 2022. Deep supervoxel segmentation for survival analysis in head and neck cancer patients, in: Lecture Notes in Computer Science (LNCS) Challenges.

Kang, L., Chen, W., Petrick, N.A., Gallas, B.D., 2015. Comparing two correlated c indices with right-censored survival outcome: a one-shot nonparametric approach. Statistics in Medicine 34, 685–703.

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., van Wijk, Y., Woodruff, H., van Soest, J., Lustberg, T., Roelofs, E., van Elmpt, W., Dekker, A., Mottaghy, F.M., Wildberger, J.E., Walsh, S., 2017. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 14, 749–762. URL: http://www.nature.com/doifinder/10.1038/nrclinonc.2017.141, doi:10.1038/nrclinonc.2017.141.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R.G.P.M., Granton, P., Zegers, C.M.L., Gillies, R.J., Boellard, R., Dekker, A., Aerts, H.J.W.L., 2012. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer 48, 441–446.

Lang, D.M., Peeken, J.C., Combs, S.E., Wilkens, J.J., Bartzsch, S., 2022. Deep learning based GTV delineation and progression free survival risk score prediction for head and neck cancer patients, in: Lecture Notes in Computer Science (LNCS) Challenges.

Lee, J., Kang, J., Shin, E.Y., Kim, R.E.Y., Lee, M., 2022. Dual-path connected CNN for tumor segmentation of combined PET-CT images and application to survival risk prediction, in: Lecture Notes in Computer Science (LNCS) Challenges.

Leger, S., Zwanenburg, A., Leger, K., Lohaus, F., Linge, A., Schreiber, A., Kalinauskaite, G., Tinhofer, I., Guberina, N., Guberina, M., et al., 2020. Comprehensive analysis of tumour sub-volumes for radiomic risk modelling in locally advanced hnscc. Cancers 12, 3047.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Liu, T., Su, Y., Zhang, J., Wei, T., Xiao, Z., 2022. 3D U-net applied to simple attention module for head and neck tumor segmentation in PET and CT images, in: Lecture Notes in Computer Science (LNCS) Challenges.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Lu, J., Lei, W., Gu, R., Wang, G., 2022. Priori and posteriori attention for generalizing head and neck tumors segmentation, in: Lecture Notes in Computer Science (LNCS) Challenges.

Ma, B., Guo, J., De Biase, A., Sourlos, N., Tang, W., van Ooijen, P., Both, S., Sijtsema, N.M., 2022. Self-supervised multi-modality image feature extraction for the progression free survival prediction in head and neck cancer, in: Lecture Notes in Computer Science (LNCS) Challenges.

Ma, J., Yang, X., 2021. Combining CNN and hybrid active contours for head and neck tumor segmentation, in: 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. Nature communications 9, 1–13.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Büttner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, A.T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kenngott, H., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K.G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., van Smeden, M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Calster, B.V., Varoquaux, G., Jäger, P.F., 2022. Metrics reloaded: Pitfalls and recommendations for image analysis validation URL: https://arxiv.org/abs/2206.01653v5.

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. Medical Image Analysis , 101796.

Martinez-Larraz, A., Asenjo, J.M., Rodríguez, B.A., 2022. PET/CT head and neck tumor segmentation and progression free survival prediction using deep and machine learning techniques, in: Lecture Notes in Computer Science (LNCS) Challenges.

Meng, M., Peng, Y., Bi, L., Kim, J., 2022. Multi-task deep learning for joint tumor segmentation and outcome prediction in head and neck cancer, in: Lecture Notes in Computer Science (LNCS) Challenges.

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 transactions on medical imaging 34, 1993–2024.

Morand, G.B., Vital, D.G., Kudura, K., Werner, J., Stoeckli, S.J., Huber, G.F., Huellner, M.W., 2018. Maximum standardized uptake value (suvmax) of primary tumor predicts occult neck metastasis in oral cancer. Scientific reports 8. doi:10.1038/S41598-018-30111-7.

Murugesan, G.K., Brunner, E., McCrumb, D., Kumar, J., VanOss, J., Moore, S., Peck, A., Chang, A., 2022. Head and neck primary tumor segmentation using deep neural networks and adaptive ensembling, in: Lecture Notes in Computer Science (LNCS) Challenges.

Naser, M.A., Wahid, K.A., van Dijk, L.V., He, R., Abobakr Abdelaal, M., Dede, C., Mohamed, A.S.R., Fuller, C.D., 2022a. Head and neck cancer primary tumor auto segmentation using model ensembling of deep learning in PET-CT images, in: Lecture Notes in Computer Science (LNCS) Challenges.

Naser, M.A., Wahid, K.A., Mohamed, A.S.R., Abdelaal Abobakr, M., He, R., Dede, C., van Dijk, L.V., Fuller, C.D., 2022b. Progression free survival prediction for head and neck cancer using deep learning based on clinical and PET-CT imaging data, in: Lecture Notes in Computer Science (LNCS) Challenges.

Oreiller, V., Andrearczyk, V., Jreige, M., Boughdad, S., Elhalawani, H., Castelli, J., Vallières, M., Zhu, S., Xie, J., Peng, Y., et al., 2022. Head and neck tumor segmentation in pet/ct: the hecktor challenge. Medical image analysis 77, 102336.

Patel, Y., Srivastava, S., Rana, D., Goel, A., Suryanarayana, K., Saini, S.K., 2021. Pet-ct scan-based maximum standardized uptake value as a prognostic predictor in oropharynx squamous cell cancer. Cancer treatment and research communications 26.

Qayyum, A., Benzinou, A., Mazher, M., Abdel-Nasser, M., Puig, D., 2022. Automatic segmentation of head and neck (H&N) primary tumors in PET and CT images using 3D-Inception-ResNet model, in: Lecture Notes in Computer Science (LNCS) Challenges.

Ren, J., Huynh, B.N., Groendahl, A.R., Tomic, O., Futsaether, C.M., Korreman, S.S., 2022. PET normalizations to improve deep learning auto-segmentation of head and neck in 3D PET/CT, in: Lecture Notes in Computer Science (LNCS) Challenges.

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.

Saeed, N., Al Majzoub, R., Sobirov, I., Yaqub, M., 2022. An ensemble approach for patient prognosis of head and neck tumor using multimodal data, in: Lecture Notes in Computer Science (LNCS) Challenges.

Salmanpour, M.R., , Hajianfar, G., Rezaeijo, S.M., Ghaemi, M., Rahmim, A., 2022. Advanced automatic segmentation of tumors and survival prediction in head and neck cancer, in: Lecture Notes in Computer Science (LNCS) Challenges.

Savjani, R.R., Lauria, M., Bose, S., Deng, J., Yuan, Y., Andrearczyk, V., 2022. Automated tumor segmentation in radiotherapy, in: Seminars in Radiation Oncology, Elsevier. pp. 319–329.

Sobirov, I., Nazarov, O., Alasmawi, H., Yaqub, M., 2022. Automatic segmentation of head and neck tumor: How powerful transformers are? arXiv preprint arXiv:2201.06251 .

Starke, S., Thalmeier, D., Steinbach, P., Piraud, M., 2022. A hybrid radiomics approach to modeling progression-free survival in head and neck cancers, in: Lecture Notes in Computer Science (LNCS) Challenges.

Vallières, 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. Scientific reports 7, 1–14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wahl, R.L., Jacene, H., Kasamon, Y., Lodge, M.A., 2009. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. Journal of nuclear medicine 50, 122S–150S.

Wang, G., Huang, Z., Shen, H., Hu, Z., 2022a. The head and neck tumor segmentation in PET/CT based on multi-channel attention network, in: Lecture Notes in Computer Science (LNCS) Challenges.

Wang, J., Peng, Y., Guo, Y., Li, D., Sun, J., 2022b. CCUT-Net: Pixel-wise global context channel attention UT-Net for head and neck tumor segmentation, in: Lecture Notes in Computer Science (LNCS) Challenges.

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 transactions on medical imaging 23, 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. Scientific Reports 2021 11:1 11, 1–15.

Wong, A.J., Kanwar, A., Mohamed, A.S., Fuller, C.D., 2016. Radiomics in head and neck cancer: from exploration to application. Translational cancer research 5, 371.

Xie, J., Peng, Y., 2021. The head and neck tumor segmentation using nnU-Net with spatial and channel 'squeeze & excitation' blocks, in: Lecture Notes in Computer Science (LNCS) Challenges.

Xie, J., Peng, Y., 2022. The head and neck tumor segmentation based on 3D U-Net, in: Lecture Notes in Computer Science (LNCS) Challenges.

Yousefirizi, F., Janzen, I., Dubljevic, N., Liu, Y.E., Hill, C., MacAulay, C., Rahmim, A., 2022. Segmentation and risk score prediction of head and neck cancers in PET/CT volumes with 3D U-Net and cox proportional hazard neural networks, in: 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, in: Lecture Notes in Computer Science (LNCS) Challenges.

Yuan, Y., Adabi, S., Wang, X., 2022. Automatic head and neck tumor segmentation and progression free survival analysis on PET/CT images, in: Lecture Notes in Computer Science (LNCS) Challenges.

Zhang, Y., Lobo-Mueller, E.M., Karanicolas, P., Gallinger, S., Haider, M.A., Khalvati, F., 2020. CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging. BMC Medical Imaging 20, 1–8.

Author contributions

Vincent Andrearczyk and Valentin Oreiller:

Design of the task and of the challenge, organization of the challenge, development of post-challenge analyses, writing of the paper.

Mario Jreige: Design of the task and of the challenge, organization of the challenge, quality control/annotations, tumors delineation.

Martin Vallières:

Design of the task and of the challenge, provided the initial data and annotations for the training set Vallières et al. (2017).

Sarah Boughdad, Catherine Cheze Le Rest, Olena Tankyevych, and Hesham Elhalawani :

Design of the task and of the challenge, tumors delineation.

John O. Prior and Dimitris Visvikis:

Design of the task and of the challenge, organization of the challenge.

Adrien Depeursinge and Mathieu Hatt:

Design of the task and of the challenge, organization of the challenge, development of post-challenge analyses, writing of the paper.

The list of scanners used in the different centers is provided in Table A.6.

Table A.6: List of scanners used in the six centers. Discovery scanners are from GE Healthcare, Biograph from Siemens, and Gemini from Phillips.

	Train					Test		
	HGJ	CHUS	HMR	CHUM	CHUP	CHUP	CHUV	Total
Discovery STE			18	56				74
Gemini GXL 16		72						72
Biograph 40					53	48		71
Discovery ST	55							55
Discovery 690							53	53

Appendix A. Scanners and Image Acquisition Information

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

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

HMR: For the PET portion of the FDG-PET/CT scan, a median of 475 MBq (range: 227-859) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 360 s (range: 120-360) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 5 (range: 3-5). The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52×3.52 mm 2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 120-140) with a median exposure of 11 mAs (range: 5-16) was used. The CT slice thickness resolution was 3.75 mm for all patients and the median in-plane resolution was 0.98×0.98 mm 2 (range: 0.98-1.37).

CHUM: For the PET portion of the FDG-PET/CT scan, a median of 315 MBq (range: 199-3182) was injected intravenously. After a 90-min uptake period of rest, patients were imaged with the PET/CT imaging system. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 120-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 3 (range: 3-5). The median FDG-PET slice thickness resolution was 4 mm (range: 3.27-4) and the median in-plane resolution was 4×4 mm 2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 120 kVp (range: 120-140) with a median exposure of 350 mAs (range: 5-350) was used. The median CT slice thickness resolution was 1.5 mm (range:

1.5-3.75) and the median in-plane resolution was 0.98 × 0.98 mm 2 (range: 0.98-1.37).

CHUV: The patients fasted at least 4h before the injection of 4 Mbq/kg of(18F)-FDG (Flucis). Blood glucose levels were checked before the injection of (18F)-FDG. If not contra-indicated, intravenous contrast agents were administered before CT scanning. After a 60-min uptake period of rest, patients were imaged with the PET/CT imaging system. First, a CT (120 kV, 80 mA, 0.8-s rotation time, slice thickness 3.75 mm) was performed from the base of the skull to the mid-thigh. PET scanning was performed immediately after acquisition of the CT. Images were acquired from the base of the skull to the mid-thigh (3 min/bed position). PET images were reconstructed by using an ordered-subset expectation maximization iterative reconstruction (OSEM) (two iterations, 28 subsets) and an iterative fully 3D (DiscoveryST). CT data were used for attenuation calculation.

CHUP: PET/CT acquisition began after 6 hours of fasting and 60±5 min after injection of 3 MBq/kg of 18F-FDG (421±98 MBq, range 220-695 MBq). Non-contrast-enhanced, non-respiratory gated (free breathing) CT images were acquired for attenuation correction (120 kVp, Care Dose® current modulation system) with an in-plane resolution of 0.853×0.853 mm2 and a 5 mm slice thickness. PET data were acquired using 2.5 minutes per bed position routine protocol and images were reconstructed using a CT-based attenuation correction and the OSEM-TrueX-TOF algorithm (with time-of-flight and spatial resolution modeling, 3 iterations and 21 subsets, 5 mm 3D Gaussian post-filtering, voxel size 4×4×4 mm3).

Appendix B. Segmentation Algorithms Summary

Table B.7 provides a synthetic comparison of the methodological choices and designs of the participants' algorithms.

			Preprocess.				Data augmentation					Model archit.				Loss				Training/ evaluation				
Team	Dice	HD95	iso-resampling	CT clipping	Min-max norm.	Standardization	Rotation	Scaling	Flipping	Noise addition	Other	U-Net	Attention	Res. connection	SE norm. Isantsen et al. (2021)	Dice	Cross-entropy	Focal Lin et al. (2017)	Else	Optimizer	nnU-Net Isensee et al. (2021)	LR decay	Cross-validation	Ensembling
Pengy	0.7785	3.0882	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓	✓		✓		SGD	✓	✓	✓	5
SJTU EIEE.[14]	0.7733	3.0882	✓	✓	✓		✓		✓			✓		✓	✓	✓	✓	✓	✓	Adam	✓	✓	✓	9
HiLab	0.7735	3.0882	✓	✓		✓						✓	✓		✓	✓		✓		Adam		✓	✓	14
BCIOQurit	0.7709	3.0882	✓	✓		✓	✓	✓	✓			✓		✓				✓	✓	Adam	✓	✓	✓	10
Aarhus Oslo	0.7790	3.1549	✓				✓	✓				✓				✓	✓			Adam	✓	✓	✓	3
Fuller MDA	0.7702	3.1432	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓				Adam		✓	✓	10
UMCG	0.7621	3.1432	✓			✓						✓	✓					✓	✓	Adam		✓	✓	5
Siat	0.7681	3.1549	✓	✓		✓	✓		✓			✓	✓	✓		✓		✓		na		na	✓	5
Heck Uihak	0.7656	3.1549	✓	✓	✓	✓	✓		✓			✓	✓		✓	✓		✓		Adam		✓	✓	5
BMIT USYD	0.7453	3.1549	✓	✓	✓	✓	✓		✓			✓				✓				Adam		✓	✓	10
DeepX	0.7602	3.2700	✓	✓		✓			✓		✓	✓	✓	✓	✓	✓				Adam		✓	✓	15
Emmanuelle B.	0.7595	3.2700	✓	✓		✓	✓		✓			✓	✓	✓	✓	✓		✓		Adam		✓	✓	5
C235	0.7565	3.2700	✓	✓	✓	✓						✓	✓	✓		✓		✓		Adam		✓		5
Abdul Qayyum	0.7487	3.2700			✓				✓		✓	✓		✓		✓	✓			Adam			✓	
RedNeucon	0.7400	3.2700	na	na	na	na	✓	✓	✓			✓	✓			✓	✓		✓	Adam		✓		25
DMLang	0.7046	4.0265	✓	✓	✓		✓		✓	✓	✓	✓				✓				Adam				
Xuefeng	0.6851	4.1932	✓	✓		✓	✓	✓	✓	✓		✓				✓	✓			SGD		✓	✓	
Qurit Tecvico	0.6771	5.4208			✓							✓				✓				Adam				
Vokyj	0.6331	6.1267	✓	✓		✓	✓									✓		✓		Adam		✓		
TECVICO Corp F.	0.6357	6.3718	na	na	na	na						✓		✓		✓				Adam				2
BAMF health	0.7795	3.0571	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓		✓	✓			SGD	✓		✓	10
Wangjiao	0.7628	3.2700	✓	✓		✓						✓	✓	✓	✓	✓		✓		Adam		✓	✓	6

Table B.7: Synthetic comparison of segmentation methods and results. The number of used models is reported in the last column when ensembling was used. "na" stands for "not available". Table reproduced from (Andrearczyk et al., 2021b)

Appendix C. Outcome Prediction Algorithms Summary

Table C.8 provides a synthetic comparison of the methodological choices and designs of participants' algorithms for tasks 2 and 3.

		Pre-processing						Segment.			Image features				Modeling and training approach											Masks		
Team	**C-index Task 2**	Iso-resampling	CT clipping	Min-max norm.	Standardization	PET/CT fusion	Further cropping	Relies on Task 1	Additional segm.	No segmentation	Deep features	Large radiomics set	Volume, shape	IBSI compliant	Ensembling	Deep model	Algo. RF, SVM...	Feature selection	PET as input	CT as input	PET/CT fusion	Use clinical var.	Imputed missing	Cross-val.	Augmentation	**C-index Task 3**	GT masks	Task 1 masks	PET thresh. masks
BioMedIA	0.7196			✓		✓	✓			✓	✓					✓					✓	✓				na			
Fuller MDA	0.6938		✓	✓	✓					✓	✓				✓	✓			✓	✓		✓			✓	0.6978	✓	✓	
Qurit Tecvico	0.6828			✓				✓				✓		✓	✓		✓	✓			✓	✓		✓		na			
BMIT_USYD	0.6710		✓	✓	✓		✓	✓			✓				✓	✓			✓	✓		✓			✓	na			
DMLang	0.6681		✓	✓				✓			✓					✓			✓	✓		✓			✓	na			
TECVICO_C.	0.6608					✓	✓	✓				✓		✓			✓	✓			✓	✓		✓		na			
BAMF Health	0.6602		✓	✓	✓			✓	✓			✓		✓			✓	✓	✓	✓		✓	✓	✓		0.6602			✓
ia-h-ai	0.6592											✓		✓			✓	✓	✓	✓		✓		✓		0.6592			✓
Neurophet	0.6495							✓					✓				✓					✓		✓		na			
UMCG	0.6445		✓	✓	✓	✓				✓	✓				✓	✓			✓	✓	✓	✓		✓	✓	0.6373	✓	✓	
Aarhus Oslo	0.6391																					✓		✓		na			
RedNeucon	0.6280							✓	✓			✓					✓	✓				✓		✓		na			
Emmanuelle B.	0.6223	✓	✓		✓			✓			✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	✓		na			
BCIOQurit	0.6116	✓	✓		✓			✓			✓	✓		✓			✓	✓	✓	✓		✓		✓	✓	0.4903	✓		
Vokyj	0.5937	✓	✓		✓			✓					✓				✓					✓		✓	✓	na			
Xuefeng	0.5510	✓	✓		✓			✓					✓				✓									0.5089	✓		
DeepX	0.5290	✓	✓		✓			✓				✓		✓			✓		✓	✓		✓				na			

Table C.8: Synthetic comparison of outcome prediction methods. All participants of task 3 also participated in task 2. Table reproduced from (Andrearczyk et al., 2021b)