Efficient Multi-Task Framework for Head and Neck Segmentation, Staging, and Prognosis

Xinlu Shi[1 [0009−0000−1286−4146]], Duy Vu-Tran[1 [0000−0003−3474−2435]], and Bulat Ibragimov[1 [0000−0001−7739−7788]]

Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
shixinlusci@gmail.com

Abstract. Head and neck cancers remain a challenging problem for computer-aided diagnosis. The HECKTOR 2026 challenge introduces a multi-modal benchmark covering three main tasks: segmentation, TN staging, and prognosis. We build one end-to-end framework where the segmentation results and clinical variables provide the input for both staging and prognosis, and evaluate it on the challenge's unseen test data. Segmentation uses nnU-Net and reaches a mean Dice of 0.694. TN staging uses multinomial logistic regression, reaching balanced accuracies of 0.661 for N stage and 0.546 for T stage. Prognosis uses an ensemble of ICARE and a random survival forest on a selected set of radiomic and clinical features, achieving a C-index of 0.639. The resulting weighted challenge score is 0.640.
Team: IMAGE.

Keywords: Head and neck cancer · PET/CT · Segmentation · TN staging · Radiomics · Survival analysis · Multi-center generalization.

1 Introduction

Head and Neck cancer is among the most frequent cancers in the world, increasing in both incidence and mortality [17]. While recent developments in medical imaging diagnosis are highly promising [18, 3, 4], their sample sizes are relatively small (around 100–400 cases), limiting their generalizability. On the other hand, balancing the number of variables and observations as well as preventing overestimation requires labor-intensive annotations on larger datasets. Thus, the HEad and neCK TumOR (HECKTOR) organizers have built a large-scale multi-center cohort that has been progressively advanced since the first edition [2, 1].

The 2026 edition extends the previous year's dataset, increasing to a total of 1423 patients and introducing a new center (UAE) [15, 16]. The dataset focuses on FDG-PET and CT scan processing, providing metabolic activity and structural information. Furthermore, the organizers encourage an end-to-end modular pipeline for HNC diagnosis spanning segmentation, TN staging, and recurrence-free survival (RFS) risk prediction. The design promotes knowledge sharing across subtasks and optimization, which reflects the real-world decision-making process. The framework involves primary tumor and lymph node recog-

nition, from which the T and N stages are classified. Finally, a survival score is regressed in the prognosis task.

Prior HECKTOR editions provide context for all three tasks. For segmentation, nnU-Net is a widely used backbone [7, 8]. For prognosis, outcome prediction has the longest track record, where radiomics-based models have remained competitive [18]. Across the 2021 and 2022 challenges, RFS or PFS C-indices remained within a narrow range (about 0.65 to 0.72), and the 2022 overview reported that doubling the training data did not materially improve the top C-index [2, 1]. The 2022 winner (ICARE) was a binary-weighted radiomics ensemble, not a deep network, and edged out a joint segmentation-and-survival network (DeepMTS [11]) by less than a thousandth of a C-index [14, 1]. Across those editions, simple models were competitive, while deep survival models showed limited gains at this sample size. TN staging is newly introduced in 2026, with reference T and N labels following the American Joint Committee on Cancer (AJCC) staging criteria [5]. The 2025 edition added lymph-node segmentation and HPV status and reported large gaps between validation and hidden-test performance on prognosis and HPV-classification [16], indicating overfitting and optimistic validation estimates.

Our work makes three main contributions. First, we present a single container that integrates all three challenge tasks (Fig. 1): an ensemble nnU-Net segmentation model, TN staging derived directly from the predicted segmentation masks, and an ensemble survival predictor. The complete pipeline operates within the computational constraint and successfully qualified for the testing phase. Second, we improve the robustness of TN staging across unseen centers by aligning training with inference. Specifically, the staging models are trained on predicted rather than ground-truth masks, and T and N categories are inferred from standardized logistic models that map mask geometry to staging labels. In contrast to the AJCC's size rule-based and tree-based classifiers, this approach avoids reliance on fixed absolute thresholds that may not generalize across imaging centers. Third, for prognosis, we prioritize cross-center generalizability by using a compact feature set comprising eight core radiomic features and two HPV-stratified tumor burden variables. These features drive an ensemble of ICARE and random survival forest (RSF) models, which demonstrates better performance than either large radiomic feature sets or deep survival models.

2 Data and Evaluation Protocol

The 2026 training release has 782 patients, each of which includes a pair of co-registered CT and PET scans, an expert-annotated primary tumor (GTVp) and lymph node (GTVn). In addition, a clinical record of each patient is provided with age, sex, tobacco and alcohol use, performance status, treatment, and HPV status. Missingness is strongly center-dependent. HPV missingness ranges from 1.6% (MDA) to 100% (USZ). RFS labels are available for 727 patients, including 148 recurrences (20%). Event prevalence also varies substantially by center (Table 1), creating center-specific patterns that models may inadvertently learn.

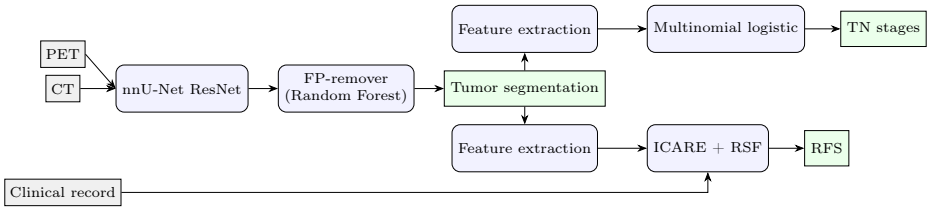


Fig. 1: Pipeline overview. The cleaned mask is the segmentation output and the shared input to TN staging and prognosis. The staging heads are trained on features from out-of-fold predicted masks, and the prognosis model is trained on features from ground-truth-masks.

Table 1: Training cohort by center. Event rate as the recurrence fraction among RFS-labelled patients.

Center	n	RFS missing	Event rate	HPV missing
CHUM	56	0.0%	12.5%	60.7%
CHUP	75	37.3%	38.3%	20.0%
CHUS	72	0.0%	18.1%	54.2%
CHUV	51	13.7%	18.2%	90.2%
MDA	444	4.5%	19.1%	1.6%
HGJ	55	0.0%	20.0%	30.9%
HMR	18	0.0%	22.2%	88.9%
USZ	11	0.0%	54.5%	100.0%
Total	782	7.0%	20.4%	23.7%

Tumor-mask composition varies across patients: 664 have both GTVp and GTVn, 87 have only the primary (all staged N0), and 31 have nodal disease with no delineated primary, so the staging model cannot assume that both structures are present. GTVn is the harder segmentation target and shows the largest variation in detection across centers. Its median volume is 13.3 ml per patient, versus 7.5 ml for GTVp. Stage labels follow the AJCC 7th edition [5], i.e., spanning T0–T4 and N0–N3, with all subcategories merged into the major one.

To estimate generalization to new centers, prognosis is evaluated with center-grouped cross-validation, where centers are assigned wholly to one of five folds so that no center appears in both training and validation within a fold. A random split lets a fold train and test on the same center, giving a number likely optimistic relative to the test set, and we quantify that gap in Section 4.4. TN staging is evaluated with leave-one-center-out (LOCO), pooling the held-out predictions across centers into a single balanced-accuracy score.

3 Methods

The container runs one segmentation, derives the TN stage and the prognosis input from the resulting mask, and writes all three outputs. The segmentation is the shared upstream artifact; TN and prognosis consume it in parallel.

3.1 Segmentation

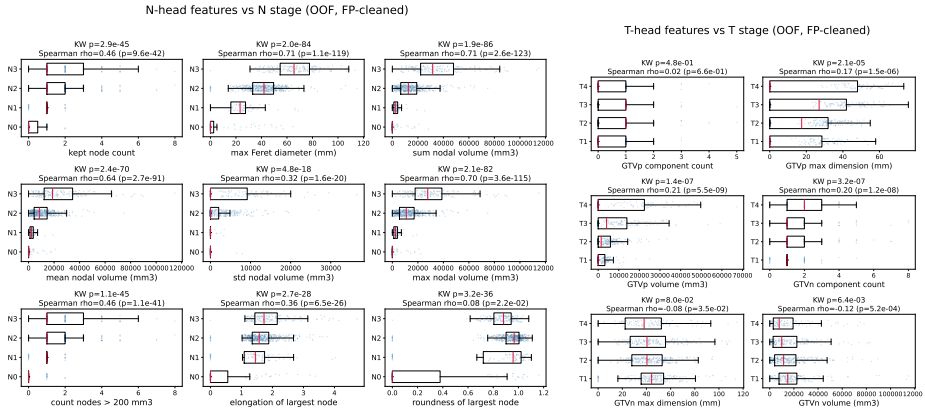
We adopted the self-configuring nnU-Net framework [7], which can automatically reconfigure the pre- and post-processing of the U-Net pipeline to achieve better results. The medium version of the ResNet encoder [8] is chosen, as it fits the challenge's compute budget while remaining effective. The model takes as input the concatenation of the registered CT and PET, resulting in a two-channel input $I \in \mathbb{R}^{2\times H\times W\times D}$. Thus, the model predicts the voxel-level object probability distribution, i.e., $\mathcal{M}(I) : \mathbb{R}^{2\times H\times W\times D} \mapsto [0,1]^{C\times H\times W\times D}$, where C is the number of object classes.

During training, the data are randomly augmented with rotation, scaling, elastic deformation, gamma, brightness, noise, and blur. The addition of Dice and Cross-entropy loss serves as our loss function. Furthermore, exponential moving average [12] is deployed to stabilize the training process. We also train the model in K-fold style and average the results with mirroring test-time augmentation and Gaussian-weighted sliding window at inference.

To address the false-positive predictions, GTVn post-processing consists of a learned node-cleanup step. For each connected component in the predicted GTVn mask, a random forest estimates the probability that the component represents a true lymph node using 19 features. These include size and shape descriptors (volume, Feret diameter, elongation, roundness, flatness, and surface-to-volume ratio), PET first-order statistics (maximum SUV, SUV standard deviation, 90^{th} percentile SUV, SUV interquartile range, peak-to-background ratio, rim-to-core ratio, and background SUV), CT first-order statistics (10th-percentile HU, low-HU fraction, and HU standard deviation), and spatial context (distance to the primary tumor, laterality, and distance to the nearest node). Components are removed when their predicted probability falls below 0.5, the classifier's default decision point. The random forest was selected using leave-one-center-out cross-validation from among gradient boosting, XGBoost, and random forest classifiers trained on the same 19-feature set, and was trained using the segmentation model's out-of-fold predictions. The resulting cleaned segmentation mask serves as the final output for Dice evaluation and as the input for downstream staging and prognostic analyses.

3.2 TN Staging

According to AJCC manual [5], the T and N stages are defined mainly by the number of nodes and the tumor's size. To verify these characteristics, we visualize the distribution of each feature across the T and N stages (Figure 2). Based



(a) (b)

Fig. 2: Relations between features and TN stages. a: N-stage statistics, b: T-stage statistics. The statistics are taken from the predicted out-of-fold (OOF) with false-positive (FP) cleaned. Kruskal-Wallis p-value and Spearman coefficient are shown on top of each plot. Red line indicates the median.

on the observed statistical significance, we deploy two multinomial logistic regression models (for N and T stages separately) using geometric and statistical features extracted from the predicted tumor masks. Table 2 summarizes the chosen features for the logistic regression of T and N staging. Both models fall back to our implementation of the AJCC size rule if feature extraction fails.

The heads are trained on the cleaned masks derived from our segmentation model's out-of-fold predictions, with the scalers fit inside the training folds only so that training matches inference. During inference, the container only sees predicted masks, which carry the false positives and fragmentation that ground-truth masks do not. On LOCO folds, we compare the standardized logistic regression heads against our implementation of the AJCC size rule and tree classifiers (Section 4.3).

3.3 Prognosis

The prognosis model uses four clinical variables (age, tobacco, treatment, HPV), eight core radiomic features, and two HPV-stratified tumor-burden interaction features. The radiomic features are the metabolic tumor volume, mean SUV and sphericity of GTVp and GTVn, the total lesion count, and the nodal count. The two interaction features enter the total metabolic tumor volume (GTVp + GTVn) separately for HPV-positive and HPV-negative patients, and are zero when HPV is missing. This gives tumor burden a distinct slope per stratum, since it appeared less prognostic in HPV-positive disease. Radiomics are extracted with PyRadiomics [6] from GTVp and GTVn at 1 mm with a fixed CT

Table 2: Features used for TN-staging logistic regression. One submitted variant adds the CT sphericity of the primary tumor for T prediction.

Task	Feature count	Features
N	9	Node count 200mm^3-and-above node count Maximum Feret diameter Nodal volume (sum, mean, std., maximum) Elongation of the largest node Roundness of the largest node
T	7	Component count of the primary tumor Component count of the lymph nodes Maximum Feret diameter of the primary tumor Maximum volume of the primary tumor Maximum Feret diameter among the lymph nodes Maximum volume of the lymph nodes Age

window and PET in SUV units, and log-transformed; clinical variables are one-hot encoded on the training fold, with missing tobacco, treatment, and HPV each assigned an "unknown" level and missing age median-imputed. The risk head is an ensemble of ICARE [14] (bagged, 100 estimators) and a random survival forest [9, 13] (300 trees, minimum leaf size 15, features per split $\sqrt{p}$). To combine them per case, each model's risk is mapped to its percentile in the stored training-fold risk distribution and the two percentiles are averaged. ICARE provides the more stable linear component as it uses sign-based standardized effects, so no single coefficient magnitude dominates.

Unlike the TN staging heads, the prognosis model is trained on features derived from the ground-truth masks over the full cohort. The container writes an RFS time, which the challenge scores anti-concordant with risk, so the combined risk is passed through a strictly decreasing map to a positive pseudo-time; only the ordering matters for the C-index. If feature extraction fails, the model falls back to a clinical-only prediction to avoid a missing score.

The features and risk head were selected based on cross-center stability. This is our central design choice, analyzed in Section 4.4.

4 Experiment and Results

The results below come from two sources: the challenge leaderboard (Section 4.1) and our center-grouped and LOCO cross-validation of the training data (Sections 4.2–4.4).

4.1 Overall setup and results

To ensure consistency across all data samples, all the CT/PET scans and ground-truth masks are resampled to a resolution of 1 mm^3 with B-spline and nearest-

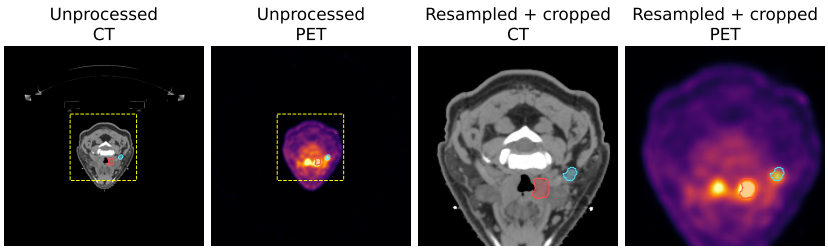


Fig. 3: Data example: from raw to processed. The dashed yellow line indicates the region of interest. Red and Cyan regions are GTVp and GTVn, respectively.

Table 3: Validation and test results. Weighted $= 0.25\,\text{Dice} + 0.35\,(\text{T} + \text{N})/2 + 0.40\,\text{C}$.

Submission	Phase	Weighted	Mean Dice	C-index	T bAcc	N bAcc
A	Val	0.6951	0.6885	0.7518	0.5025	0.7673
A	Test	0.6374	0.6939	0.6389	0.5005	0.6900
B	Test	0.6403	0.6939	0.6389	0.5458	0.6613

neighbour interpolation, respectively. Furthermore, head-and-neck regions are cropped using the distribution of the PET's SUV, which allows the models to attend to corresponding regions. The chosen crop size is $200 \times 200 \times 310\text{mm}^3$, which covers the head and neck of all patients within the training set. Figure 3 shows an example of the image preprocessing. Feature computation for TN staging and prognosis is based on the $1\,\text{mm}^3$ prediction.

GTVp and GTVn predictions are evaluated with mean Dice score, with the GTVn measured by average object-based Dice. TN staging is measured with average class-wise balanced accuracy, while C-index is the metric for prognosis.

Table 3 presents the leaderboard scores of our two submissions. Compared to Submission A, Submission B adds a CT-sphericity feature to the T head and raises the nodal false-positive remover's threshold from 0.5 to 0.6, which improved N in our LOCO evaluation. Submission B was built after the validation window closed so it has no validation score. On the test set, B reaches higher T and lower N, and gives the higher weighted score. The test scores are about 0.06 below the validation score, with the drop mainly concentrated in prognosis and N while segmentation and T stay relatively stable.

4.2 Segmentation

We train the model using stochastic gradient descent for 300 epochs with a batch size of 2 and a patch size of $112 \times 112 \times 192\text{mm}^3$. The learning rate decays polynomially from 0.01. The foreground sampling probability is set to 0.5, forcing the model to see and recognize more lesion regions. We also normalize the SUV with Z-score, while the CT scans are scaled by the nnU-Net framework.

Table 4: TN staging, pooled balanced accuracy on leave-one-center-out folds over out-of-fold predicted masks.

Task	AJCC / size rule	Standardized LR
N	0.446	**0.626**
T	0.364	**0.546**

The four-fold ensemble reaches a mean Dice of about 0.68 on our out-of-fold local validation. GTVp achieves higher segmentation accuracy than GTVn (out-of-fold Dice approximately 0.74 vs. 0.62). The nodal errors mainly consist of false-positive components: spurious nodes, over-segmentation, and fragmentation of single lesions into several components. We therefore apply a learned false-positive remover to the GTVn mask. However, large false-positive components often overlap true nodes in size and uptake, limiting the effectiveness of post-processing. Further improvement likely requires a more node-sensitive segmentation model.

4.3 TN Staging

On LOCO folds over the out-of-fold masks, we compare our multinomial logistic regression models against our implementation of the AJCC size rule, which is also the pipeline's fallback when feature extraction fails (Table 4). Tree classifiers on the same features stalled between 0.42 and 0.51, suggesting weaker cross-center transfer than the logistic heads, possibly because the trees' fixed split thresholds were less robust to center-specific feature shifts.

The weaker T head reflects the mask-derived geometry separating the T stages far less cleanly than the N stages. As shown in Figure 2, the T stages differ little across the AJCC features (high p-values), leaving the model little signal to separate them, whereas the N-stage features show a clear separation ($p < 0.05$), which explains the stronger N head. Even when trained and validated on ground-truth masks, LOCO T balanced accuracy stays around 0.5 across logistic and tree-based classifiers, indicating that the T ceiling is not driven by segmentation error. N-stage errors are dominated by upstream GTVn failures. On ground-truth masks, our AJCC rule implementation reaches around 0.98 N balanced accuracy under LOCO, indicating that most of the predicted-mask N gap is due to segmentation quality. Centers with low GTVn Dice also have the lowest N balanced accuracy. Further N improvement is therefore likely to require better nodal detection in the segmentation model.

4.4 Prognosis and Center-Aware Evaluation

Table 5 reports C-index under random and center-grouped cross-validation across feature sets and model families. The eight-feature core was selected from a larger feature pool by keeping features with consistent prognostic value under

Table 5: Prognosis C-index by feature set, random vs. center-grouped 5-fold CV. core8 is the eight-feature core and HPV-strat the two HPV-stratified tumor-burden features, both defined in Section 3.3. † marks the deployed configuration.

Features	Model	Random CV	Grouped CV
Clinical-4	ICARE	0.620	0.582
+ full pool	ICARE	0.609	0.566
+ core8	ICARE	0.657	0.637
+ ground-truth T/N	ICARE	0.641	0.612
+ core8 + ground-truth T/N	ICARE	0.658	0.640
+ core8	ICARE+RSF	**0.670**	0.657
+ core8 + HPV-strat†	ICARE+RSF	0.667	**0.662**

the center-grouped protocol. The full pool comprised 400 PyRadiomics features (first-order, shape, and texture over CT and PET for GTVp and GTVn) and 11 tumor-burden features (metabolic tumor volume, SUV statistics, and lesion counts). For the radiomics features, the eight-feature core provided incremental value over the clinical variables, improving the grouped ICARE C-index from 0.582 to 0.637, whereas inclusion of the full radiomics feature set degraded performance, indicating feature dilution. During model selection, ICARE and the RSF achieved the highest performance. Penalized Cox models (ridge and LASSO) and gradient-boosted Cox models performed within one fold standard deviation of the top models, while the deep learning approaches (DeepSurv [10] and MTLR [19]) consistently underperformed. An ensemble combining the two best-performing models outperformed each individual model.

In a paired comparison, the ICARE C-index dropped by only 0.01 when ground-truth masks were replaced with out-of-fold predicted masks. Mean SUV and metabolic tumor volume are strongly correlated between the two mask types (Spearman correlation: 0.84–0.90), whereas the count features are not (0.19–0.34), with the predicted masks containing roughly twice as many components due to fragmented lesions and false-positive nodes. The modest performance degradation suggests that the more consistent features mitigate the impact of segmentation errors.

5 Discussion and Conclusion

For prognosis, the center-grouped C-index (0.66) is close to the test C-index (0.64), while the validation-leaderboard C-index (0.75) is about 0.11 higher. The leaderboard validation is computed on a relatively small sample (about 50 cases), resulting in substantial estimation variance. Random splitting provides better stability but can introduce center-level data leakage and lead to optimistic performance estimates. The center-grouped cross-validation approach uses the complete training cohort while enforcing center-level separation between folds, thereby mitigating both the variance and the center-level bias.

The primary factors constraining overall pipeline performance are the limited sample size and the accuracy of the segmentation models. Prognosis is bounded by the low event count (148), while the two smallest centers (USZ n=11, HMR n=18) introduce noise. Incorporating ground-truth T and N staging variables does not improve the grouped C-index, suggesting that staging information provides limited additional prognostic value beyond the tumor-burden features in the current feature set, among which metabolic tumor volume and lesion counts carry most of the signal. The clearest opportunity for improvement is nodal segmentation. N accuracy is gated by GTVn detection, which degrades substantially at some training centers, notably CHUP, where post-processing could not reliably distinguish large false-positive components from true nodes. A more reliable node-sensitive segmentation would improve both Dice and N. T staging is the weakest component, and unlike N it does not improve on ground-truth masks, indicating that its ceiling comes from the features rather than from segmentation quality. Our current features describe the size and shape of the tumor, whereas the T4 stage is defined by invasion into adjacent structures. Invasion is also hard to read from the dataset's PET/CT, which lacks the contrast enhancement that would better resolve the boundary between the tumor and adjacent structures [15].

We presented a single-container pipeline for segmentation, TN staging, and prognosis in HECKTOR 2026, reaching a weighted test score of 0.6403. Segmentation is a four-fold nnU-Net ensemble whose masks feed both staging and prognosis. Standardized logistic regression models predict the T and N stages from mask geometry and age, and an ensemble of ICARE and a random survival forest predicts the risk score from a compact feature set.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andrearczyk, V., Oreiller, V., Abobakr, M., Akhavanallaf, A., Balermpas, P., Boughdad, S., Capriotti, L., Castelli, J., Cheze Le Rest, C., Decazes, P., Correia, R., El-Habashy, D., Elhalawani, H., Fuller, C.D., Jreige, M., Khamis, Y., La Greca, A., Mohamed, A., Naser, M., Prior, J.O., Ruan, S., Tanadini-Lang, S., Tankyevych, O., Salimi, Y., Vallières, M., Vera, P., Visvikis, D., Wahid, K., Zaidi, H., Hatt, M., Depeursinge, A.: Overview of the hecktor challenge at miccai 2022: Automatic head and neck tumor segmentation and outcome prediction in pet/ct. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 1–30. Springer Nature Switzerland, Cham (2023)
2. Andrearczyk, V., Oreiller, V., Boughdad, S., Rest, C.C.L., Elhalawani, H., Jreige, M., Prior, J.O., Vallières, M., Visvikis, D., Hatt, M., Depeursinge, A.: Overview of the hecktor challenge at miccai 2021: Automatic head and neck tumor segmentation and outcome prediction in pet/ct images. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 1–37. Springer International Publishing, Cham (2022)

3. Bogowicz, M., Riesterer, O., Stark, L.S., Studer, G., Unkelbach, J., Guckenberger, M., Tanadini-Lang, S.: Comparison of pet and ct radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. Acta Oncologica **56**(11), 1531–1536 (Aug 2017)
4. Castelli, J., Depeursinge, A., Ndoh, V., Prior, J., Ozsahin, M., Devillers, A., Bouchaab, H., Chajon, E., de Crevoisier, R., Scher, N., Jegoux, F., Laguerre, B., De Bari, B., Bourhis, J.: A pet-based nomogram for oropharyngeal cancers. European Journal of Cancer **75**, 222–230 (2017)
5. Edge, S.B., Compton, C.C.: The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. Annals of Surgical Oncology **17**(6), 1471–1474 (Feb 2010)
6. van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21), e104–e107 (Oct 2017)
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Dec 2020)
8. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jäger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 488–498. Springer Nature Switzerland, Cham (2024)
9. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. The Annals of Applied Statistics **2**(3) (2008)
10. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Research Methodology **18**(1), 24 (2018)
11. Meng, M., Gu, B., Bi, L., Song, S., Feng, D.D., Kim, J.: Deepmts: Deep multitask learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment pet/ct. IEEE Journal of Biomedical and Health Informatics **26**(9), 4497–4507 (2022)
12. Morales-Brotons, D., Vogels, T., Hendrikx, H.: Exponential moving average of weights in deep learning: Dynamics and benefits (2024)
13. Pölsterl, S.: scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research **21**(212), 1–6 (2020)
14. Rebaud, L., Escobar, T., Khalid, F., Girum, K., Buvat, I.: Simplicity is all you need: Out-of-the-box nnunet followed by binary-weighted radiomic model for segmentation and outcome prediction in head and neck pet/ct. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 121–134. Springer Nature Switzerland, Cham (2023)
15. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., Khan, U., Ridzuan, M., Andrearczyk, V., Depeursinge, A., Xie, Y., Eugene, T., Metz, R., Dore, M., Delpon, G., Papineni, V.R.K., Wahid, K., Dede, C., Ali, A.M.S., Sjogreen, C., Naser, M., Fuller, C.D., Oreiller, V., Jreige, M., Prior, J.O., Rest, C.C.L., Tankyevych, O., Decazes, P., Ruan, S., Tanadini-Lang, S., Vallières, M., Elhalawani, H., Abgral, R., Floch, R., Kerleguer, K., Schick, U., Mauguen, M., Bourhis, D., Leclere, J.C., Sambourg, A., Rahmim, A., Hatt, M., Yaqub, M.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction (2025)

16. Saeed, N., Hassan, S., Hardan, S., Cai, L., Liang, X., Mazher, M., Qayyum, A., Bu, Y., Lyu, M., Lin, Y., Meng, M., Huang, C., Wang, L., Chamseddine, D., Ahrari, S., Wu, B., Chen, Y., Mao, F., Zhang, H., Zhao, B., Ray, S., Guo, M., Xiang, L., Dexl, J., Ingrisch, M., Depeursinge, A., Rahmim, A., Hatt, M., Andrearczyk, V., Yaqub, M.: Head and neck tumor (hecktor) 2025: Benchmark of segmentation, diagnosis, and prognosis in multimodal pet/ct (2026)
17. Sun, H., Yu, M., An, Z., Liang, F., Sun, B., Liu, Y., Zhang, S.: Global burden of head and neck cancer: Epidemiological transitions, inequities, and projections to 2050. Frontiers in Oncology **Volume 15 - 2025** (2025)
18. Vallières, M., Kay-Rivest, E., Perrin, L.J., Liem, X., Furstoss, C., Aerts, H.J.W.L., Khaouam, N., Nguyen-Tan, P.F., Wang, C.S., Sultanem, K., Seuntjens, J., El Naqa, I.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific Reports **7**(1) (Aug 2017)
19. Yu, C.N., Greiner, R., Lin, H.C., Baracos, V.: Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In: Advances in Neural Information Processing Systems. vol. 24, pp. 1845–1853 (2011)