Oral Oncology 144 (2023) 106460

Contents lists available at ScienceDirect

Oral Oncology

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

Spatially-aware clustering improves AJCC-8 risk stratification performance in oropharyngeal carcinomas

Guadalupe Canahuate [a,*], Andrew Wentzel [b], Abdallah S.R. Mohamed [c], Lisanne V. van Dijk [c], David M. Vock [d], Baher Elgohari [c], Hesham Elhalawani [c], Clifton D. Fuller [c], G. Elisabeta Marai [b]

[a] *Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA*
[b] *Department of Computer Science, The University of Illinois at Chicago, Chicago, IL 60612, USA*
[c] *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA*
[d] *Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA*

ARTICLE INFO

Keywords:
Head & Neck cancer
Statistical and research methods
Machine learning
Radiation oncology
Health Informatics
Imaging
Risk stratification
Oropharyngeal cancer
Clustering
Radiomics

ABSTRACT

Objective: Evaluate the effectiveness of machine learning tools that incorporate spatial information such as disease location and lymph node metastatic patterns-of-spread, for prediction of survival and toxicity in HPV+ oropharyngeal cancer (OPC).
Materials & methods: 675 HPV+ OPC patients that were treated at MD Anderson Cancer Center between 2005 and 2013 with curative intent IMRT were retrospectively collected under IRB approval. Risk stratifications incorporating patient radiometric data and lymph node metastasis patterns via an anatomically-adjacent representation with hierarchical clustering were identified. These clusterings were combined into a 3-level patient stratification and included along with other known clinical features in a Cox model for predicting survival outcomes, and logistic regression for toxicity, using independent subsets for training and validation.
Results: Four groups were identified and combined into a 3-level stratification. The inclusion of patient stratifications in predictive models for 5-yr Overall survival (OS), 5-year recurrence free survival, (RFS) and Radiation-associated dysphagia (RAD) consistently improved model performance measured using the area under the curve (AUC). Test set AUC improvements over models with clinical covariates, was 9 % for predicting OS, and 18 % for predicting RFS, and 7 % for predicting RAD. For models with both clinical and AJCC covariates, AUC improvement was 7 %, 9 %, and 2 % for OS, RFS, and RAD, respectively.
Conclusion: Including data-driven patient stratifications considerably improve prognosis for survival and toxicity outcomes over the performance achieved by clinical staging and clinical covariates alone. These stratifications generalize well to across cohorts, and sufficient information for reproducing these clusters is included.

Introduction

Head and neck cancers (HNCs) affect almost 65,000 individuals per year in the United States, with approximately 14,000 deaths from the disease [1]. The prognosis of HNCs is considerably variable in different tumor types, ranging from excellent prognosis, as in Human papillomavirus (HPV)-associated squamous cell carcinoma [2,3], to deadly disease as in advanced HPV-negative tumors [4,5]. The incidence of oropharyngeal cancer (OPC) has been increasing for the last few decades. The increased prevalence of HPV-positive cases has also led to improved treatment outcomes and has motivated the modification of the AJCC staging system and TNM Classification of Malignant Tumors (TNM), which is a standardized system for classifying the spread and extent of cancer for use in treatment planning and as a prognostic tool [6]. The current staging system relies however on only the primary tumor's size and extension, and size and laterality of secondary nodal tumors, while overlooking other relevant features such as radiomics or the disease spread [6].

The ability to better personalize treatment approaches and further treatment efficacy requires better risk stratification models so that patients with lower risk may benefit from treatment de-escalation (i.e. reduction of long-term toxicity) whereas patients with higher risk may benefit from treatment intensification strategies (i.e. increased tumor control rate) [7,8].

[*] Corresponding author.
E-mail address: guadalupe-canahuate@uiowa.edu (G. Canahuate).

https://doi.org/10.1016/j.oraloncology.2023.106460
Received 8 September 2021; Received in revised form 26 May 2023; Accepted 5 June 2023
Available online 28 June 2023

Imaging radiomics is a method that extracts a large number of features from patients' images. These features can identify tumor characteristics that cannot be appreciated by the naked eye. The inclusion of imaging radiomics in risk stratification of cancer patients showed promising results for many cancer sites [9–11]. In addition to radiomics, we recently showed that the anatomically-informed clustering of the lymph node patterns-of-spread (LN) is associated with treatment outcomes [12].

To this end, we sought to evaluate the effect of including patient risk stratifications derived from radiomics and patterns of lymph node metastasis to improve the prediction of oncologic and toxicity outcomes in a large cohort of oropharyngeal cancer patients. We consider two survival outcomes: Overall Survival (OS) and Recurrence-Free Survival (RFS), and radiation associated dysphagia (RAD) as a toxicity outcome associated with OPC patients.

Methods

Data

Patients were retrieved from an internal University of Texas MD Anderson Cancer Center (UT MDACC) database after approval from the UT MDACC Institutional review board (IRB). All methods for this study were performed in accordance with the UT MDACC IRB guidelines and regulations.

Our original cohort consists of 575 patients that were randomly split into two independent datasets for training (N = 391) and validation (N = 284) before the start of the study. Inclusion criteria for this study where: 1) histopathologically-proven squamous cell carcinoma of the OPC; 2) tumor present at the base of tongue, tonsil, soft palate, pharyngeal wall, glossopharyngeal sulcus, or vallecula; 3) HPV/p16 positive status assessed via in-situ hybridization or immunohistochemistry; 4) available pre-treatment contrast-enhanced CT scans, with contours for the primary gross tumor volume (GTVp); and 5) patients were treated with curative-intent intensity-modulated radiation therapy with concurrent chemotherapy.

Clinical features including age at diagnosis, sex, ethnicity, smoking status and frequency, subsite of the primary tumor within the oropharynx, T category, N category, therapeutic combination, and AJCC stage (7th and 8th edition) were extracted from electronic medical records. A detailed description of these data can be found in Elhalawani et al. [13].

We consider two survival outcomes and one toxicity outcome. Overall survival (OS) refers to the number of months survived after diagnosis or last follow-up time (for censored outcomes). Recurrence Free Survival (RFS) is a combined survival outcome including Local Control, Regional Control, and Distant Metastasis, whichever occurs first, or last follow-up time (for censored outcomes). RAD was defined as the presence of grade 2+ aspiration rate based on CTCAE guidelines [18], or feeding-tube insertion during treatment or after treatment completion [37]. No feeding tubes were placed prophylactically.

For imaging data, contrast-enhanced computed tomography (CECT) scans acquired at diagnosis, prior to any local or systemic treatment, were exported via commercially available contouring software (Velocity AI v3.0.1). 3D volumes of interest (VOIs) including the gross primary tumor volumes (GTVp) were segmented by a radiation oncologist, and then inspected by a second radiation oncologist. The generated VOIs and CT images were exported to DICOM-RTSTRUCT format to be used for radiomics features extraction. The primary tumor volumes (GTVp) were contoured based on the ICRU 62/83 definition [14] and radiomics features representing intensity, shape, and texture were extracted using the freely available open-source software IBEX [15].

For radiomics, we extracted thousands of human-defined and curated features which describe tumor shape, intensity, and texture, among other characteristics [43]. This enabled us to choose and engineer radiomic features proven to be more immune to inter-scanner variability, boosting generalizability and significant clinical correlativity [44]. Acknowledging concerns for inter- and intra-observer variability associated with manual segmentations, we assigned two expert radiation oncologists who were blinded to relevant clinical data. Discrepancies were resolved by consensus or the call of a third expert radiation oncologist. To decrease volume-dependence of radiomic features, pixels were resampled to 1 mm × 1 mm and Laplacian of gaussian and Butterworth smoothing were applied to non-shape feature extraction with standard deviations between 0.5 and 2.5 [45]. Previous work from our group has found that inter-observer variability of the selected radiomics features is low relative to inter-patient variability in squamous cell carcinoma [42,46,47]. The small number of features that fell below the acceptable stability threshold were excluded in this analysis.

Regions in the lymph node drainage system (levels) that were affected (nodal tumors) were annotated for each patient. Involved lymph nodes in each patient were identified for both sides of the head. If at least one lymph node in a given level is affected with cancer cells, radiation oncologists refer to the corresponding node level as being involved with disease, and they involve the whole node level in treatment. Involvement was treated as separate covariates for the side of the head with the primary tumor (ipsilateral) and the side opposite the primary tumor (contralateral). When the primary tumor (GTVp) crossed the midline of the head (bilateral), the side with the larger bulk of primary disease was treated as the ipsilateral side. A multidimensional vector was constructed to describe the patterns of affected nodes in a way that accounted for the relative anatomical positions of the lymph node levels, which was used for creating clusters of patients based on similar LN involvement, as described by Luciani et al.[16].

Data preprocessing

Some radiomic features that relied on larger filters or large neighborhood sizes could not be extracted from smaller tumor volumes. In these cases, missing radiomics were imputed using Multivariate Imputation by Chained Equations [17], using classification and regression trees (CART). Imputation of the training data was done first, and then the validation samples were imputed using the complete training data. No clinical data was imputed, and patients with missing data or unknown HPV status were excluded from the analysis.

A 5-year cutoff was used to generate an event indicator for each survival outcome. Only patients that experienced the outcome before the 60-month mark were flagged as having experienced the event.

Radiomics clusters and features

We created clusters of patients based on radiomic features as follows. First, a set of 3831 features were extracted using the IBEX package. Features with zero variance or that were highly correlated (>80 %), and then centered and scaled using the Caret R package [19]. Additionally, based on previous studies using the same cohort that identified tumor volume and intensity as highly predictive for local control [20], the features F25.ShapeVolume (first-order feature), and F29.IntesityDirectGlobalMean (shape feature) are always included a-priori in the final set of features. Using the final set of features, a penalized semi-parametric Cox regression model [21], which was tuned using cross-validation, was applied to select the most informative radiomic features. A radiomic signature was generated with the selected features and the linear predictor from the Cox model was used as a radiomic score. Lloyds (K-means) clustering was applied to generate patient stratification with three groups based on radiomic information, referred to as RM clusters from here onwards.

Lymph-node similarity and clustering

Pairwise similarity between patients was calculated using similarity-based on lymph node (LN) involvement over adjacent anatomical regions as described by Wentzel et al [12]. Similarity was computed using the squared Canberra distance metric [22] using an anatomically-aware encoding of patients based on the patterns of involved LN levels. Hierarchical clustering with four clusters was performed using Ward's linkage method [23] on a subset of the patient in the training set that had involved lymph nodes, based on the parameters used in the original study [12]. These four clusters were created from the LN spread patterns of the patients in the training cohort. Patients in the validation cohort were assigned to the clusters with their corresponding pattern. Patients in the validation cohort with patterns not present in the original cohort were assigned to the nearest existing cluster based on average euclidean distance, while patients with no lymph node involvement were grouped into their own fifth cluster. Finally, the five LN clusters were grouped into the low-risk group (No involvement or cluster 1), and a high-risk group (clusters 2–4), based on the relative incidence of toxicity found in the training cohort. For the remainder of this paper, "LN clusters" will refer to these two (high or low) risk strata, rather than the original five clusters.

The RM and LN clusters were then combined into a RLN stratification with three risk groups. RLN 1 is a low risk group and corresponds to the patients in low risk groups for both RM and LN. RLN 2 is a medium risk group for patients with low/medium risk for either RM or LN, and RLN 3 is a high-risk group with patients in either high risk for RM or LN.

Statistical analysis

Kaplan-Meier curves for OS and RFS were computed for strata defined by AJCC stage (8th edition) and RLN precision-imaging stratification. We compared OS and RFS among these strata using the log rank test. We then assessed the improvement of including the cluster labels as covariates in a Cox proportional hazards model, and evaluated the prediction improvement over the same baseline model without the cluster labels, e.g. using only the clinical features and/or clinical staging. Clinical covariates included age at diagnosis, smoking status (current/former/never) and whether the patient received chemotherapy or not (yes/no). Clinical staging covariates include AJCC (8th edition). The a-priori tumor volume and intensity radiomic features (2F) were also tested as predictive covariates for some of the models. Models were built over the training dataset and evaluated over the validation dataset.

Several metrics are used to evaluate the results. Over the training data we compute the Akaike information criterion (AIC) as a measure of the goodness of fit and simplicity of the model. We evaluate the improvement in model discrimination and calibration when including the radiomics and LN cluster labels for the validation datasets. For model discrimination, we computed both the area under the curve (AUC) of the Receiver Operating Characteristic (ROC), which considers sensitivity against specificity for consecutive cutoffs of the survival probability, and Harrel's C-index (i.e. probability of concordance). For toxicity outcomes we only report AUC, as the C-index is identical to AUC for binary outcomes [24]. For evaluating model calibration, we computed Brier score and the Nam-D'Agostino test statistic [25], which are suggested as relevant metrics in the literature [26]. All statistical analysis was performed using statistical software R version 3.2.3.

Results

Table 1 shows the patient demographics for clinical features, as well as the precision-imaging patient stratifications and survival outcomes considered. As expected, both training and validation sets follow the same demographic distribution. For the two survival outcomes, about 16 % of the patients experienced the event before the 5-year cutoff for OS and RFS. For the toxicity outcome, 23 % of the patients experienced dysphagia.

The Coxnet model selected 9 radiomic features using cross-validation over the training data (Table 2). Four radiomics-derived clusters were identified to represent different risk groups. The low and medium risk groups account for 95 % of the patients, while the high-risk group is the smallest group (5 % of patients) (Table 1).

Table 1
Summary of clinical and demographic features, follow-up time and event rate for the outcomes considered, as well as the data-driven patient stratifications for both training and validation sets. Table shows median (25th, 75th percentile) for continuous values and count frequency (%) for discrete values.

Number of patients	Training		Validation	
	391		284	
Covariates	Median and 25th-75th percentile or Count Frequency (%)		Median and 25th-75th percentile or Count Frequency (%)	
Age	58.04	52.25–65.32	58.15	53.38–64.10
Gender				
Male	346	88 %	243	86 %
Female	45	12 %	41	14 %
T Category				
T1/T2	281	72 %	182	64 %
T3/T4	110	28 %	102	36 %
N Category (8th ed)				
N0/N1	302	77 %	196	69 %
N2/N3	89	23 %	88	31 %
AJCC Stage (8th ed)				
I	239	61 %	143	50 %
II	100	26 %	86	30 %
III	52	13 %	55	20 %
IV	0	0 %	0	0 %
Smoking Status				
Former	146	37 %	109	38 %
Current	80	20 %	34	12 %
Never	165	42 %	141	50 %
Tumor subsite				
Tonsil	159	41 %	126	44 %
Base of Tongue	195	50 %	140	49 %
Other	37	9 %	18	6 %
Therapeutic Combination				
CC	203	52 %	76	27 %
IC + CC	91	23 %	110	39 %
IC + Radiation Alone	30	8 %	62	22 %
Radiation Alone	67	17 %	36	13 %
Response				
Overall Survival (OS)				
Alive	327	84 %	241	85 %
Deceased	64	16 %	43	15 %
Survival/Follow-up Time (in months)	56.80	43.83–80.95	59.60	45.65–70.95
Relapse Free Survival (RFS)				
Alive	320	82 %	244	86 %
Deceased	71	18 %	40	14 %
Survival/Follow-up Time (in months)	52.53	39.57–77.57	58.45	42.15–69.33
Dysphagia (RAD)				
Yes	79	20 %	79	28 %
No	312	80 %	205	72 %
Patient Stratifications				
RLN Clusters				
1	108	20 %	64	12 %
2	250	64 %	177	62 %
3	33	8 %	43	15 %

For the LN clusters, the low-risk group corresponds to patients with no lymph node involvement and cluster 1 (78 % of the training data and 72 % of the validation), while the remaining groups are considered medium to high-risk. Fig. A1 in Appendix A shows a summary of the LN clusters found in the training data.

To aid interpretation, Fig. 1 shows a visual summary [34] of the combined RLN clustering for the entire cohort of patients. Whereas AJCC is an aggregate risk staging system, our model adds additional anatomical information, and whereas our model correlates with AJCC, it has additional capacity at a more granular risk prediction.

Fig. 2 shows the Kaplan-Meier (KM) curves for OS over training and validation stratified by AJCC staging, and the combined radiomics and LN clustering (RLN). There are significant differences in OS among the strata in the training set curves ($p < .001$), as well as the validation curves ($p < .01$).

Table 2
Radiomic features: name, mean, standard deviation, and weight coefficient from the Coxnet model, as well as the linear predictor cutoffs to determine the cluster labels, description and number missing from each dataset.

Feature Name	Description	Feature Type	Mean ()	Std Dev ()	Weight (w)	Missing (Training)	Missing (Validation)
F8.IntensityDirectKurtosis	Measures the peakedness of all the voxel intensities.	First-Order	41.81	54.05	0.012	86	0
F12.IntensityDirectGlobalMean	Mean intensity along all voxels	First-Order	1.06	0.09	0.006	86	4
F13.IntensityDirectEnergy	Magnitude of voxel values	First-Order	255.92	824.15	0.001	86	20
F25.ShapeVolume	The physical volume of the voxels	Shape	12.25	14.61	0.207	86	0
F28.GrayLevelRunLengthMatrix25…0LongRunLowGrayLevelEmpha	The joint distribution of long-run lengths with lower gray-level values	Gray Level Run Length Matrix (GLRLM)	0.00	0.02	0.017	86	0
F29.IntensityDirectGlobalMin	Minimum intensity along all voxels	First-Order	296.22	259.89	−0.207	86	0
F39.NeighborIntensityDifference25Contrast	Measure of the spatial intensity change, dependent on the overall gray-level dynamic range.	Neighboring Gray Tone Difference Matrix (NGTDM)	0.32	1.48	0.059	86	0
F52.NeighborIntensityDifference25Complexity	Measure of the uniformity and number of rapid changes in gray level intensity	Neighboring Gray Tone Difference Matrix (NGTDM)	410056.46	1559470.22	0.149	86	0
F29.IntensityDirectLocalRangeMax	Maximum of the neighborhood intensity range of each voxel	First-Order	1223.34	313.45	0.116	0	0

[a]Definition of the radiomics-derived clusters. The linear predictor is calculated as $LP = \sum_{\{i=1-9\}} w_i \frac{F_i - u_i}{\sigma_i}$. The cutoffs to determine the cluster labels are $[-\infty, -0.197, 0.213, 0.866, \infty]$ and correspond to the midpoint between the cluster centroids.

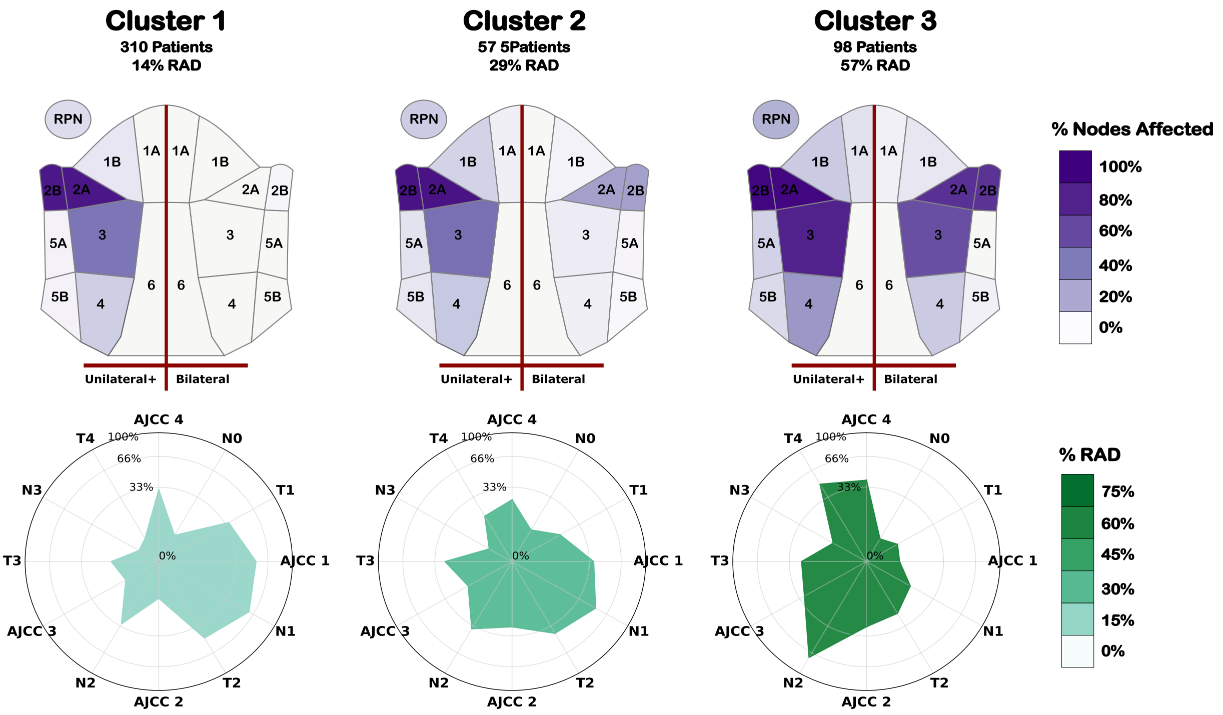

Fig. 1. Average lymph node involvement and clinical staging categories of the 3 derived lymph node + radiomics clusters across both the training and validation datasets. (Top) Heat map of the percentage of patients with an involved lymph node within a cluster for each level. The left half of each heatmap encodes patients with at least one node involved, whereas the right encodes patients with bilateral involvement in the given level. The low-risk cluster has no bilateral nodal involvement, while the highest risk cluster has significantly higher bilateral involvement and disease spread in levels 3, 4 and 5. (Bottom) Radar chart showing the % of patients within each cluster with a given staging level. The plot shading is mapped to the incidence of late treatment associated with dysphagia. The low-risk cluster is predominantly N-stage 1 and T stage 1–2, while the high-risk cluster is predominantly N-stage 2 with higher incidence of T-stage 4 and AJCC-stage 3–4.

Fig. 3 shows the corresponding Kaplan-Meier (KM) curves for RFS. Training curves show significant differences for training ($p < .001$) and validation ($p < .01$). For clustering, the validation curves show the same behavior as the training curves which is a good indicator that predictive models built over the training data generalize well to unseen data.

Table 3 shows the performance of the prediction models when RLN

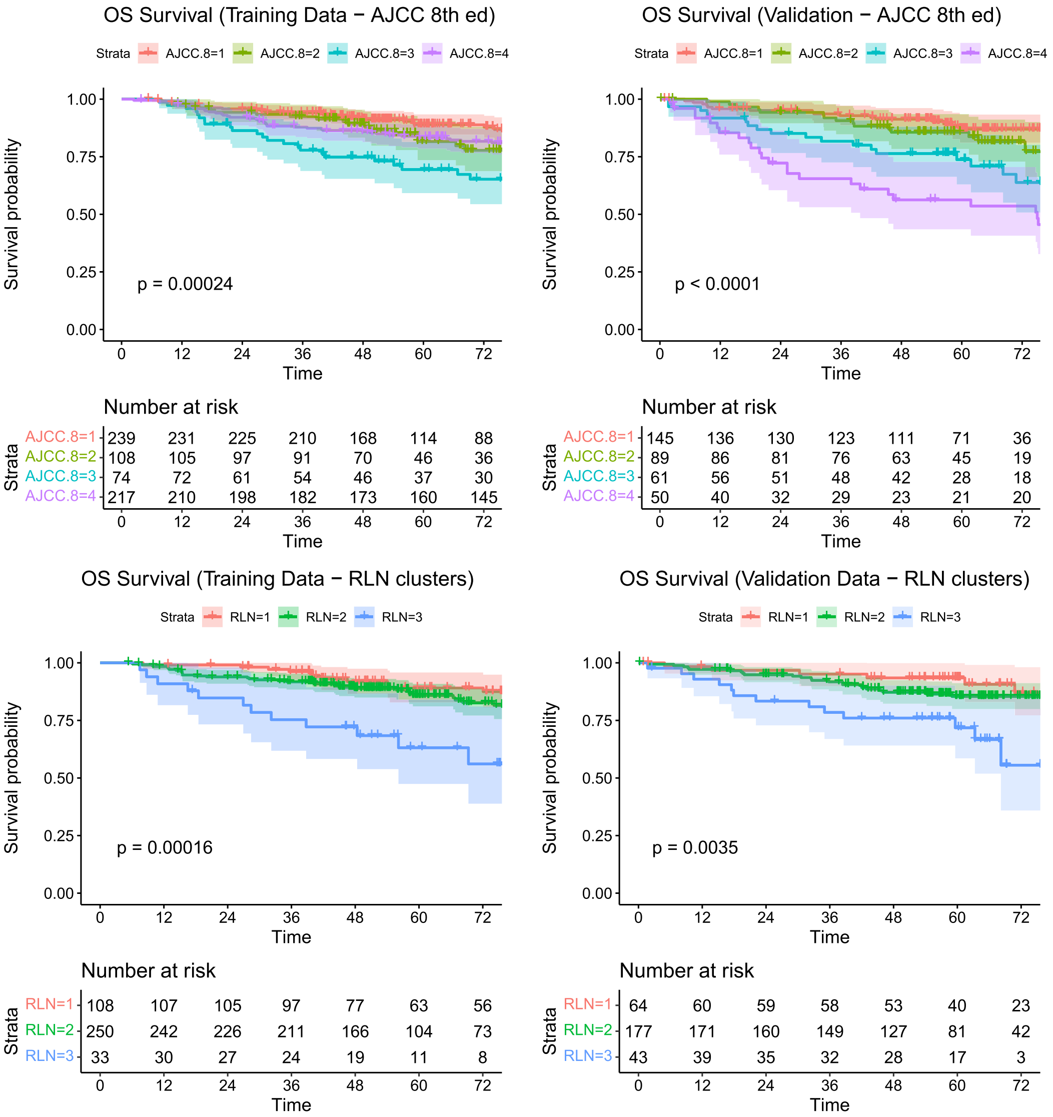

Fig. 2. Kaplan-Meier curves for Overall Survival (OS) over the training (left) and validation datasets (right), stratified by AJCC staging (2.a and 2.b), respectively stratified by the combined RLN clusters (2.c and 2.d). The RNL clusters show a significantly improved separation between the curves (p-val < 0.01), and yield the same stratification for both training and validation datasets.

cluster labels are also included in the model. A CoxPh model was trained for survival outcomes (OS and RFS) over AJCC staging (8th edition) alone and also over relevant clinical covariates (age, smoking status, and chemo). A logistic regression model was trained for dysphagia (RAD) as a toxicity outcome. AIC is reported for training, while C-index and AUC are reported for validation as measures of discrimination. The Nam-D'Agostino test statistic over the validation set is used as a measure of model calibration. Brier scores were consistently between 0.11 and 0.13 for all models and are not included in the table for conciseness. Performance of the models on the training dataset, as well as baseline performance of clusters or AJCC staging alone are included in the Appendix (Table A1).

The performance of the Baseline model is considerably improved when including the precision-imaging clusters RLN in the model (Table 3). Models that include AJCC and clinical factors outperform models with clinical factors alone for all outcomes. In terms of AIC and validation AUC, modes are improved through the addition of the combined Radiomics + Lymph node (RLN) clusters. AUC improves by 7 % (0.61–0.65) for OS, 8.52 % (0.61 vs 0.66) for RFS, and 1.6 % (0.74 vs 0.71) for Dysphagia. All improvements are larger when considering clinical-only models, possibly due to the overlap between the correlation between AJCC staging and lymph node spread, and tumor shape captured by the RLN clusters.

The best performing models for each outcome are highlighted in red in Table 3. Clinical + AJCC models with combined RLN clusters and fixed radiomic features (2F) performed the best in terms of validation AUC for OS and Dysphagia, while Clinical + AJCC + RLN performed the best for RFS. Radiomics clusters alone improved model predictions for OS and Dysphagia, while LN clusters alone improved model predictions for RFS. Models with AJCC included outperformed the baseline model with only clinical attributes.

Discussion

Our findings demonstrate that the simultaneous inclusion of covariates derived from imaging radiomics and anatomical patterns of lymph node metastasis improves the prediction of both toxicity and oncologic outcomes when compared to the standard of care staging system. The models that include Radiomics + LN consistently have superior discrimination and calibration compared to models that do not include

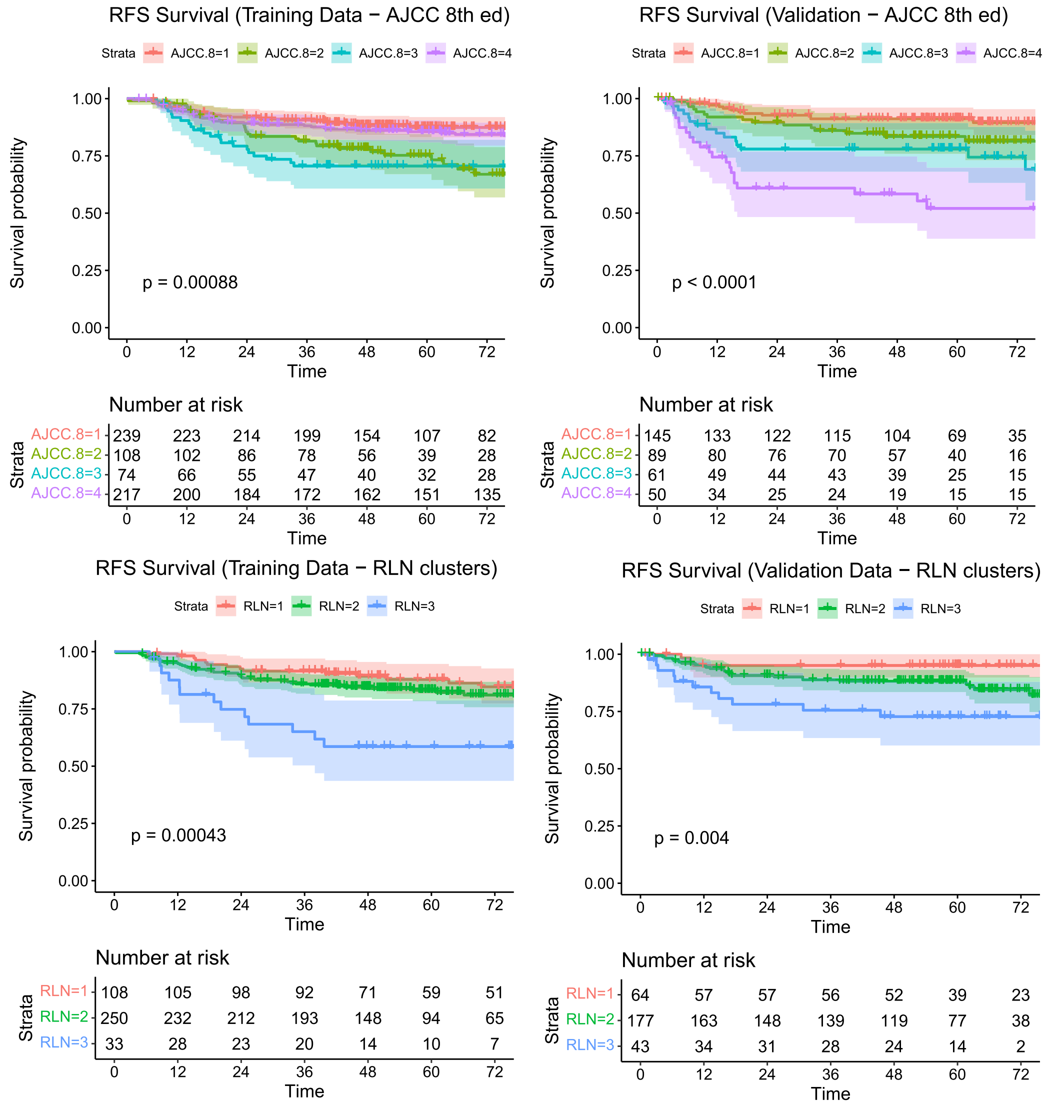

Fig. 3. Kaplan-meier curves for Recurrence Free Survival (RFS) over the training (left) and validation datasets (right), stratified by AJCC staging (3.a and 3.b), and combined RLN clusters (3.c and 3.d). The RNL clusters show a significant separation between the curves (p-val < 0.01), and the same stratification for both training and validation datasets.

these features. The improvement in discrimination over the hold-out test set indicates that the proposed patient stratifications generalize well, and offer predictive ability for both toxicity and oncologic outcomes even when the model includes other proven predictive covariates.

The KM curves stratified by the RLN clusters show significant differences in the expected survival outcomes for both training and validation (p-val < 0.01). Moreover, the curves show the same relative stratification, i.e. patients in cluster 1 show better prognosis than patients in cluster 3 for both training and testing. While the KM curves for OS and RFS (Figs. 1 and 2) stratified by AJCC staging are also significant, the training set for RFS shows an inversion for AJCC staging (8th edition), where stage III has a better expected survival than stage II. This can be partially explained by the fact that AJCC staging is optimized for OS. In contrast, the RLN clusters correlate well with both survival outcomes, OS and RFS, as well as RAD, a toxicity outcome.

Several recent studies have demonstrated correlation of radiomic features of the primary tumor and of the lymph nodes with toxicity [27–28] and oncologic [29–32] outcomes. However, these studies use the radiomic features directly into the models, and as shown in a large study, the reproducibility and robustness of these models trained using radiomic signatures for predicting OS are not warranted [38]. In contrast, we use the radiomic features to identify a discrete variable, i.e. the cluster label, to be subsequently used as a predictive covariate, and we account for anatomical LN patterns of spread. Our experiments consistently show that the cluster label as a risk strata offers better generalization than the raw radiomic features [35]. Finally, prior studies focus on the improvement of a single toxicity or oncologic outcome while in this work we propose the same stratification for improving both toxicity and oncologic outcomes simultaneously.

Our study, however, has some limitations. First, all analyses were done using a single institution retrospective dataset and an independent validation dataset from the same institution. The performance of our risk prediction models was not prospectively evaluated. Finally, given the good prognosis for oropharyngeal cancer, there is a relatively small number of events (i.e. failure and death) in the data that may introduce uncertainty in the results. In the future, we would like to include radiomic features from lymph nodes [41] as well as anatomical information relating to tumor location and organs at risk [33,36] into the patient stratifications as well as other end-points and toxicity outcomes [39,40].

In conclusion, our results demonstrate that precision risk stratifications derived from imaging data can improve the performance of

Table 3
Effect of including RLN clusters in a CoxPh model for prediction of Overall Survival (OS) and Recurrence Free Survival (RFS) at 5 years, and Radiation Associated Dysphagia (RAD) at 6 months on the validation data. Clinical features are age, smoking status (never/current/ former), and chemo (yes/no). (RM) radiomics clusters alone, (LN) Lymph node clusters, (RM + LN) combined 3-stage radiomics and lymph node clusters, (2F) a-priori radiomics features. The best and worst model in terms of validation AUC score are shown in green and red, respectively. Combined radiomics and LN clusters performed the best in all cases.

Model		Overall Survival (OS)				Recurrence Free Survival (RFS)				Dysphagia (RAD)
		AIC	Calib	Cind	AUC	AIC	Calib	Cind	AUC	AUC
Clinical + AJCC	None (Baseline)	696.75	14.34	0.64	0.61	797.55	19.47	0.62	0.61	0.73
	Additional Covariates	**ΔAIC**	**ΔCalib**	**ΔCind**	**ΔAUC**	**ΔAIC**	**ΔCalib**	**ΔCind**	**ΔAUC**	**ΔAUC**
	+RM+LN	-4.016	-5.414	0.026	0.043	**-2.921**	32.18**9**	0.04**1**	**0.052**	0.012
	+RM+LN+2FA	-5.207	-0.617	0.037	**0.060**	-1.981	23.511	0.030	0.041	**0.017**
	+2F	-4.750	-6.946	0.019	0.034	-1.687	-1.572	-0.004	-0.002	0.005
Clinical		AIC	Calib	Cind	AUC	AIC	Calib	Cind	AUC	AUC
	None (Baseline)	705.07	9.91	0.61	0.60	799.42	30.07	0.54	0.55	0.65
	Additional Covariates	**ΔAIC**	**ΔCalib**	**ΔCind**	**ΔAUC**	**ΔAIC**	**ΔCalib**	**ΔCind**	**ΔAUC**	**ΔAUC**
	+RM+LN	-9.385	-6.322	0.045	0.059	-6.126	-0.822	0.097	**0.099**	0.044
	+RM+LN+2FA	-14.386	-2.920	0.061	**0.070**	-5.398	-0.933	0.096	0.091	**0.080**
	+2F	-13.704	1.556	0.045	0.045	-4.219	6.262	0.049	0.040	0.058

[a] AJCC is AJCC (8th edition).
[b] Clinical covariates used are age, smoking status (never/current/ former), chemo (yes/no).
[c] 2F includes tumor volume and intensity as predictive radiomic features directly into the model.

predictive models for oropharyngeal cancer patients' toxicity and oncologic outcomes. The proposed stratifications incorporate anatomical information available at diagnosis such as radiomic features of the primary tumor, as well as patterns of lymph node spread. In our analyses, the addition of the cluster labels as predictive covariates consistently improves model AUC performance when compared to the same models only including clinical covariates and cancer staging. The performance improvement over the hold out test set shows the models are generalizable to previously unseen data for both oncologic outcomes such as OS and RFS as well as toxicity outcomes such as radiation associated dysphagia.

Declaration of Competing Interest

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

Data availability

Compliant with our NIH funding, we will make a public access anonymized dataset available for fair reuse. The data will be embargoed until publication, after which it will be available at https://doi.org/10.6084/m9.figshare.23154254.

Acknowledgements and sources of support

This work was supported by the National Institutes of Health (NIH) Big Data to Knowledge (BD2K) Program of the National Cancer Institute (NCI) Early-Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (R01CA214825), and from the QuBBD program of the National Cancer Institute (R01CA225190).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.oraloncology.2023.106460.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA: Cancer J Clin 2019;69 (1):7–34.
[2] Fakhry C, Westra WH, Li S, Cmelak A, Ridge JA, Pinto H, et al. Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial. J Natl Cancer Inst 2008;100(4):261–9.
[3] Ang KK, Sturgis EM. Human papillomavirus as a marker of the natural history and response to therapy of head and neck squamous cell carcinoma. In: Seminars in radiation oncology, Vol. 22. Elsevier; 2012. p. 128–42.
[4] Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med 2010;363(1):24–35.
[5] Dahlstrom KR, Calzada G, Hanby JD, Garden AS, Glisson BS, Li G, et al. An evolution in demographics, treatment, and outcomes of oropharyngeal cancer at a major cancer center: a staging system in need of repair. Cancer 2013;119(1):81–9.

[6] O`Sullivan B, Huang SH, Su J, Garden AS, Sturgis EM, Dahlstrom K, et al. Development and validation of a staging system for hpv-related oropharyngeal cancer by the international collaboration on oropharyngeal cancer network for staging (icon-s): a multicentre cohort study. Lancet Oncol 2016;17(4):440–51.

[7] Forner D, Rigby MH, Wilke D, Taylor SM, Lamond N. Risk stratification models in human papillomavirus-associated oropharyngeal squamous cell carcinoma: the Nova Scotia distribution. J Otolaryngol-Head Neck Surg 2019;48(1):3.

[8] Bigelow EO, Seiwert TY, Fakhry C. Deintensification of treatment for human papillomavirus-related oropharyngeal cancer: current state and future directions. Oral Oncol 2020;105:104652.

[9] Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (i or ii) non—small cell lung cancer. Radiology 2016;281(3):947–57.

[10] Bera K, Velcheti V, Madabhushi A. Novel quantitative imaging for predicting response to therapy: techniques and clinical applications. Am Soc Clin Oncol Educ Book 2018;38:1008–18.

[11] Parmar C, Leijenaar RT, Grossmann P, Velazquez ER, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Sci Rep 2015;5:11044.

[12] Wentzel A, Luciani T, van Dijk L, Elgohari B, Mohamed AS, Canahuate G, et al. Precision association of lymphatic disease spread with radiation-associated toxicity in oropharyngeal squamous carcinomas. medRxiv 2020.

[13] Elhalawani H, Mohamed AS, White AL, Zafereo J, Wong AJ, Berends JE, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Sci Data 2017;4:170077.

[14] J. of the International Commission on Radiation Units and Measurements. 4. Definition of volumes. J Int Commiss Radiat Units Measur 2010;10(1):41–53.

[15] Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. ibex: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys 2015;42:1341–53.

[16] Luciani T, Wentzel A, Elgohari B, Elhalawani H, Mohamed A, Canahuate G, et al. A spatial neighborhood methodology for computing and analyzing lymph node carcinoma similarity in precision medicine. J Biomed Inform 2020;X.

[17] van Buuren S, Groothuis-Oudshoorn C. mice: multivariate imputation by chained equations in r. J Stat Softw 2011;45(3). Open Access.

[18] U. D. of Health, H. Services et al. Common terminology criteria for adverse events (ctcae) version 4.0. National Institutes of Health, National Cancer Institute; 2009.

[19] Kuhn M, et al. Building predictive models in r using the caret package. J Stat Softw 2008;28(5):1–26.

[20] M. D. A. C. C. Head, N. Q. I. W. Group, et al. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. Sci Rep 2018;8.

[21] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 2011;39(5):1.

[22] Jurman G, Riccadonna S, Visintainer R, Furlanello C. Canberra distance on ranked lists. In: Proceedings of advances in ranking NIPS 09 workshop, Citeseer; 2009.

[23] Strauss T, von Maltitz MJ. Generalising ward's method for use with Manhattan distances. PLoS ONE 2017.

[24] Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015.

[25] Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biom J 2006;48(6):1029–40.

[26] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12).

[27] Pota M, et al. Early prediction of radiotherapy-induced parotid shrinkage and toxicity based on CT radiomics and fuzzy classification. Artif Intell Med 2017;81:41–53.

[28] van Dijk LV, et al. Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. Sci Rep 2019;9(1):1–8.

[29] Wu J, et al. Integrating tumor and nodal imaging characteristics at baseline and mid-treatment computed tomography scans to predict distant metastasis in oropharyngeal cancer treated with concurrent chemoradiotherapy. Int J Radiat Oncol Biol Phys 2019;104(4):942–52.

[30] Zhai T-T, et al. Improving the prediction of overall survival for head and neck cancer patients using image biomarkers in combination with clinical parameters. Radiother Oncol 2017;124(2):256–62.

[31] Vallieres M, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep 2017;7(1):1–14.

[32] Zdilar L, et al. Evaluating the effect of right-censored end point transformation for radiomic feature selection of data from patients with oropharyngeal cancer. JCO Clin Cancer Inform 2018;2:1–19.

[33] Wentzel A, et al. Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. Radiother Oncol 2020;148:245–51.

[34] Wentzel A et al. Explainable spatial clustering: leveraging spatial data in radiation oncology. In: IEEE VIS 2020, arXiv preprint arXiv:2008.11282; 2020.

[35] Tosado J, et al. Clustering of largely right-censored oropharyngeal head and neck cancer patients for discriminative groupings to improve outcome prediction. Sci Rep 2020;10(1):1–14.

[36] Wentzel A, et al. Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration. IEEE Trans Vis and Comp Graph 2019;26(1):949–59.

[37] Christopherson KM, et al. Chronic radiation-associated dysphagia in oropharyngeal cancer survivors: towards age-adjusted dose constraints for deglutitive muscles. Clin Transl Radiat Oncol 2019;18:16–22.

[38] Ger RB, et al. Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT-and PET-imaged head and neck cancer patients. PLoS ONE 2019;14(9):e0222509.

[39] Marai GE, et al. Precision risk analysis of cancer therapy with interactive nomograms and survival plots. IEEE Trans Vis Comp Graph 2018;25(4):1732–45.

[40] Sheu T, et al. Conditional survival analysis of patients with locally advanced laryngeal cancer: construction of a dynamic risk model and clinical nomogram. Sci Rep 2017;7:43928.

[41] Elhalawani H, et al. Machine learning applications in head and neck radiation oncology: lessons from open-source radiomics challenges. Front Oncol 2018;8:294.

[42] Ger RB, Zhou S, Chi PCM, et al. Comprehensive investigation on controlling for CT imaging variabilities in radiomics studies. Sci Rep 2018;8:13047. https://doi.org/10.1038/s41598-018-31509-z.

[43] Afshar P, Mohammadi A, Plataniotis KN, Oikonomou A, Benali H. From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. IEEE Signal Process Mag 2019;36(4):132–60. https://doi.org/10.1109/MSP.2019.2900993.

[44] Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring CT scanner variability of radiomics features. Invest Radiol 2015;50(11):757.

[45] Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl Cancer Res 2016;5(4):349–63.

[46] Liu R, Elhalawani H, Mohamed AS, Elgohari B, Court L, Zhu H, et al. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. Clin Transl Radiat Oncol 2020;1(21):11–8.

[47] Korte JC, Cardenas C, Hardcastle N, Kron T, Wang J, Bahig H, et al. Radiomics feature stability of open-source software evaluated on apparent diffusion coefficient maps in head and neck cancer. Sci Rep 2021;11(1):17633.