Accepted manuscript at https://doi.org/10.1016/j.radonc.2023.109638

Prognostic radiomic signature for locally advanced head and neck cancer: development and validation on a multi-centric MRI dataset.

Abstract

Background and purpose Prognosis in locally advanced head and neck cancer (HNC) is currently based on TNM staging system and tumor subsite. However, quantitative imaging features (i.e., radiomic features) from magnetic resonance imaging (MRI) may provide additional prognostic info. The aim of this work is to develop and validate an MRI-based prognostic radiomic signature for locally advanced HNC.

Materials and methods Radiomic features were extracted from T1- and T1-weighted MRI (T1w and T2w) using the segmentation of the primary tumor as mask. In total 1072 features (536 per image type) were extracted for each tumor. A retrospective multi-centric dataset (n=285) was used for features selection and model training. The selected features were used to fit a Cox proportional hazard regression model for overall survival (OS) that outputs the radiomic signature. The signature was then validated on a prospective multi-centric dataset (n=234). Prognostic performance was evaluated using C-index. Additional prognostic value of the radiomic signature was explored.

Results: The radiomic signature had C-index=0.64 in the validation set. The addition of the radiomic signature to other clinical features (TNM staging and tumor subsite) increased prognostic ability for both OS (HPV- C-index 0.63 to 0.65; HPV+ C-index 0.75 to 0.80) and DFS (HPV- C-index 0.58 to 0.61; HPV+ C-index 0.64 to 0.65).

Conclusion: An MRI-based prognostic radiomic signature was developed and prospectively validated. Such signature can successfully integrate clinical factors in both HPV+ and HPV- tumors.

Accepted manuscript at https://doi.org/10.1016/j.radonc.2023.109638

Introduction

Head and neck squamous cell carcinoma (HNC) accounts for 3% of cancers worldwide (1). The majority presents with locally advanced disease which is amenable to curative treatments that include a combination of surgery, chemotherapy and radiotherapy. One of the most relevant prognostic factors is tumor stage. Modern imaging like contrast enhanced magnetic resonance imaging (MRI), computerized tomography (CT) or positron emission tomography (PET) are recommended for disease staging. However, tumor biology may affect prognosis, making the staging-based forecast of the single patient often not reliable (2). In this scenario, additional biomarkers are needed to refine stage-based decision making. Quantitative high throughput analysis of medical imaging, i.e. radiomics, is a potential source for those biomarkers. In this context, the development of prognostic models for overall survival (OS) and disease-free survival (DFS) has been explored the most (3, 4). Despite MRI being of the most informative imaging modalities for HNC, most radiomics studies were based on either CT or PET (3–7), while only a minority of the studies is based on MRI (8, 9). This is likely due to : i) MRI was introduced later in the context of HNC imaging; ii) MRI acquisitions have some intrinsic issues that may limit quantitative analysis (10, 11). Such intrinsic issues include the MRI signal variability of the image acquisition parameters, like resolution, time of repetition and echo and magnetic field strength, but also to inter-scanner variability within and between vendors.

The purpose of our study was to develop a prognostic MRI-based radiomic signature from a retrospective multi-centric and multi-subsite locally advanced HNC dataset with images obtained using different scanners and acquisition parameters, and to validate its prognostic performance on a prospective cohort.

Materials and methods

Patient population

Patients included in this study were part of the BD2Decide project (NCT02832102), a multi-centric clinical study in locally advanced HNC (stage III-IVa/b according to TNM 7th edition) (12). Patients were recruited in 7 cancer centers across 3 countries (Italy, Germany and the Netherlands). Follow-up data (including death and cancer recurrences) were recorded for each patient. The protocols were approved by the Ethical Committees of the participating centers, and data acquisition followed the General Data Protection Regulation of the EU.

The BD2Decide database comprised two subsets: a retrospective (1086 patients), and prospective one (451 subjects), that served as training and validation set, respectively.

The patient population included tumors from oropharynx, larynx, hypopharynx and oral cavity. Tumor staging was assessed according to both TNM 7th and 8th editions, but only the latter was used. HPV testing was mandatory in oropharyngeal carcinomas specimens (13). Standard testing included p16 immuno-histochemistry, followed, in case of p16 positivity, by an HPV DNA test (in situ hybridization and/or HPV DNA polymerase chain reaction). Non-oropharyngeal carcinomas were considered as HPV- (12).

Study inclusion criteria were: 1) at least 2 years of follow-up for alive patients; 2) availability of at least one spin-echo MR image sequence (either T1- or T2-weighted); 3) images acquired with 1.5 T scanner; 4) absence of magnetic artifacts or movement-related artifacts in the MRI images; 5) visibility of the primary tumor in the study MRI. Figure 1a-b shows the consort diagram. Different Treatment type (combination of radiotherapy, surgery and chemotherapy) was collected but not included in the models. No study defined criteria were set for MRI imaging acquisition parameters.

Clinical endpoints

The study objective was to assess the independent prognostic performance of MRI radiomic model over the 8th edition staging system in locally advanced HNC. Primary endpoint was OS, secondary

endpoint DFS. OS was defined as the time between the diagnosis of primary HNC and the date of death or last follow-up. DFS was defined as the time between the diagnosis of primary HNC and the date of first recurrence or death. Exploratory endpoints were the association of the radiomic model with predefined clinical variables (stage, subsite, HPV), and the assessment of the prognostic independency of the radiomic model over a clinical model that included stage, subsite and HPV

Image acquisition

T1-weighted (T1w) and T2-weighted (T2w) images were acquired using turbo spin-echo pulse sequences. There were variations in image acquisition parameters: time of repetition (TR), time of echo (TE), pixel spacing (PS) and slice thickness (ST). Additional information about the MR image acquisition parameters grouped by center and cohort (training/validation) are reported in supplementary material A (Tables S1-S5).

Image segmentation

Image contouring of the gross tumor volume (GTV) was performed at participating center by using the same software based on Coupled Shape Modeling (CoSMo) for image semi-automatic segmentation with the possibility of manual correction (14). The region of interest (ROI), corresponding to the primary tumor, was performed by HNC expert radiologists and it was performed slice by slice to produce a3D ROI. Only one ROI was drawn per tumor and was segmented using the T2w sequence as a reference, when available. After T2w segmentation, lesion margins were checked and corrected based on other imaging sequences (pre-contrast T1w or, when available, post-contrast T1w or diffusion-weighted images). ROIs included necrotic areas.

Image preprocessing

To account for the different MRI volume acquisition parameters and the presence of intensity modifications due to speckle noise and local variation of the magnetic field different steps of image preprocessing were applied (15). First, a 3D Gaussian filter with a 3x3x3 voxel kernel and σ=0.5 was used to denoise the images. Then, the N4ITK algorithm (16) was used to estimate and correct intensity non-uniformities due to local variations of the magnetic field. Intensity standardization in each volume was performed using Z-score (removing the mean and dividing by the standard deviation of the grey values) to ensure that each MRI image had similar ranges of signals. Voxel size resampling to an isotropic resolution of 2 mm (as in (17)) was performed with B-spline interpolation. Image preprocessing was performed in Matlab (v2019a).

Radiomic features extraction

Extraction was performed using Pyradiomics 3.0 (18). A total of 1072 radiomic features, 536 per image type (T1w and T2w) were extracted. The 536 features belonged to different categories: shape and size (14 features), first order statistics (18 features), textural (40 features), wavelet (464 features). Textural features were computed using the grey level co-occurrence matrix (GLCM) and the grey level run length matrix (GLRLM). For the wavelet features, we extracted first order and textural from the 8 first order wavelet transforms that are obtainable by applying combinations of low and high pass filters to the original 3D volumes. (LLL, LLH, LHL, LHH, HLL, HLH, HHL, HHH, were H and L represent high and low pass filter in the i-th direction). Coiflet 1 wavelets (the default of Pyradiomics) were used for the wavelet decomposition. The full list of radiomic features is available in Pyradiomics documentation (19). A fixed-bin histogram discretization (32 bins) was used prior to features extraction.

Radiomic features processing and radiomic signature training

Going from the radiomic features to the final signature is a multi-step process which is totally performed in Matlab 2019a. The whole process in in accordance with the TRIPOD checklist, which is also provided in the supplementary material B.

The first postprocessing step was Z-score normalization, to ensure that the different features had similar ranges of values. The mean and standard deviation used for the feature-wise normalization were estimated from the training set. The same parameters were saved to be used also in the validation set.

The second step was missing data imputation, since for a minority of the patients (27 retrospective and 9 prospective) one of the two MRI sequences (either T1w or T2w) was not available. The imputation was performed independently for each feature. For each feature, the missing values are imputed by random sampling from a normal distribution (gaussian with 0 mean and 1 standard deviation). We preferred this methodology compared to median/mean imputation to include an element of variability in the imputed values.

Optimal feature selection was based on: 1) stability; 2) non-redundancy; 3) prognostic significance in both univariate and multivariate Cox analyses

To select a stable radiomic feature set, we considered: variability due to changes in image acquisition parameters, simulating acquisition heterogeneity among different centers, and variability due to changes in the ROI, used to mimic the effect of inter-reader variability in the segmentation. Stability of the features to these two sources of variability were evaluated as described in (15, 20). Features were considered stable if the intra-class correlation coefficient (ICC) was above 0.75.

Redundant features were removed using pairwise correlation. When a pair of features had an absolute Spearman correlation coefficient above 0.85 only one of the two was kept. In particular, the one with lower mean correlation with all the other *n-2* features was selected.

The third step of feature selection is supervised (i.e., it uses the information about survival times and cesnsoring of the training set) and it allows to select the optimal features set to go along with volume.

This is done throughout 100 bootstrap iterations of the original training set. In each bootstrap sample the following steps are performed:

- Removal of volume which is used as baseline feature.
- Series of Univariate Cox Analysis to filter features significantly associated with survival (coefficients with $p<0.05$ after Benjamini-Hoechberg correction)
- Sorting of the selected features by C-index (descending order)
- Forward addition of features to volume into a multivariate cox model. This is done until the model keeps increasing the validation C-index (computed on the cases left out of bootstrap sampling) or it reaches the maximum number of features (which is the number of uncensored data in the training set divided by 10).
- Collection of the remaining features

The 100 bootstrap iterations provide a ranking of the features based on number of times they have been selected. The N_{opt} most frequently selected features are added, together with volume, in a final multivariate cox model which is retrained on the entire training set. The output of this model is the radiomic signature. N_{opt} is computed as the median of the number of the selected features in the different bootstrap iterations. The median value of the signature in the training set is used as the threshold to dichotomize the patients in high risk (above threshold) and low risk (below threshold). All these steps are summarized in Figure S8.

Statistical analysis

The radiomic signature independence of the vendor was evaluated using a 2-way ANOVA, including tumor site as a factor to correct for their non-uniform distribution across the different centers (Table S7). To maximize the sample number in each subgroup, training and validation datasets were merged. Independence analysis was performed separately on dead and alive patients at last follow-up.

The radiomic signature prognostic performance was evaluated on both the training and validation sets based on: Harrel's C-index (21) ; hazard ratio (HR) ; p-value of the log-rank test (22) comparing the Kaplan-Meier curves (23) for high and low risk groups. The performance of the signature was compared to the one of volume alone. 95% confidence interval are computed using bootstrap.

The correlations of the radiomic signature with tumor subsite, HPV and stage were also evaluated. Statistically significant associations were identified using Spearman correlation coefficient and/or Kruskal-Wallis test. The analysis of correlation to anatomical site and HPV was performed on all the 519 patients (training and validation set merged). In addition, since staging is not equivalent for HPV- and HPV+ tumors, correlation with stage was analyzed for HPV+ patients (n=148) and HPV- patients (n=371) separately.

For all the aforementioned statistical test, $p<0.05$ was used as a threshold of significance and the Tukey's honestly significant difference procedure (the default for Matlab 2019a) was used to perform multiple comparisons, when necessary.

Moreover, we compared the C-indexes of purely clinical vs combined clinical-radiomic models. Two clinical models were defined: TNM stage (8th edition for HPV+ tumors); subsite and TNM stage (8th edition for HPV- tumors), The HPV+ model did not need any training since it was based on one clinical variable only. The HPV- model was obtained by training a Cox proportional hazard regression analysis. Dummy data were used to represent categorical variables.

We also computed the C-indexes of clinical and corresponding clinical and radiomic model on 100 different sets obtained via bootstrap. We used as quality metrics indicating improvement the number of times the combined model performs better than the corresponding clinical model. Moreover, we computed the following metric:

$$t = \frac{mean(C_{i,comb} - C_{i,clin})}{std(C_{i,comb} - C_{i,clin})}$$

Where $C_{i,clin}$ and $C_{i,comb}$ are the C-indexes of the clinical and combined model obtained on the i-th dataset. We indicated this metric with t because it is similar to the t-statistics used in the paired t-test

but uses the standard deviation of the differences instead of the standard error of the mean, which would cause a biased metric, since the standard error of the mean tends to 0 when the bootstrap iterations are increased.

Moreover, we compared the Kaplan-Meier curves obtained stratifying by radiomic signature in accordance with stage, tumor subsite and HPV status. This latter analysis is available in the supplementary material A (Figures S19-S29).

Results

In total, 285 and 234 patients constituted the training and validation set, respectively. Clinical data for the selected patients are displayed in Table 1. Of the 519 patients, 148 were p16+ and HPV+. Therapeutic strategies used in the BD2Decide patient cohort have been previously detailed. (12). Median follow-up was 63.38 months (95% confidence interval 60.06-65.42) for the retrospective series and 27.66 months (95% CI 26.48-28.32) for the prospective one. In the overall patient cohort (519 cases), 5-year OS was 66.5%, and 5-year DFS 54.9%.

The list of stable features included 222 features (T1w: 67, T2w: 155). After correlation-based feature selection, this number was reduced to 38 (T1w: 12; T2w: 26). The full lists are available in the Supplementary material C.

The prognostic signature trained on the retrospective cohort was composed of 5 features listed in Table 2: volume and 4 wavelet features (1 related to texture and 3 related to first order statistics). Among the four non-volumetric features, 1 was extracted from T1w and 3 were extracted from T2w images. Means and standard deviations used for the Z-score normalization are reported in Table S6. Figure S9 shows the correlation matrix of the 5 features. The pairwise Spearman correlation among features ranges between -0.12 and 0.69. Therefore, we can say that the correlation-based features selection step managed to avoid redundancy. The p-value for the global Schonefeld residual is 0.35 for OS and 0.93 in DFS, so we can say that the requirements of the Cox model are met.

The validation C-indexes of the model based on tumor volume was 0.58 [95% CI: 0.50-0.67] and 0.57 [95% CI: 0.50-0.62] for OS and DFS respectively, while the corresponding C-indexes based on radiomic signatures are 0.64 [95% CI: 0.58-0.70] and 0.60 [95% CI: 0.54-0.66], and equivalent results are displayed in the supplementary material A. Figure 2 shows the Kaplan-Meier curves for the dichotomized risk groups for OS and DFS. For both models, the median value of the signature/volume in the training set is used to dichotomize the risk groups in the validation set. The equivalent results for the training set are displayed in the supplementary material A (Figure S10). The results shown are referred to the signature trained using both T1w and T2w features. Results obtained using the entire feature selection and training pipeline but with only T1w features and only T2w features are displayed in the supplementary material A (Figure S11-S12, Table S7-S8). However, single sequence-based models were had inferior prognostic performance compared to the model trained using both sequences (C-index 0.61 and 0.59 in the validation for T1w- and T2w-based models respectively).

Figure 3 illustrates the different distributions of the radiomic signature according to the clinical characteristics for the merged dataset of 519 patients.

Figure 3a-b shows the distribution of the radiomic signature by stage in HPV- and HPV+ tumors, respectively. The signature values were significantly different across stages in both groups. In HPV- carcinomas, stage IV had significantly higher signature compared to stage III ($p=0.03$). In HPV+ carcinomas, the signature increased with stage ($p=0.01$) and the difference between stage I and III was statistically significant ($p=0.01$). Figure 3c shows the distribution of the radiomic signature by tumor subsite. There were significant differences among the distributions of the radiomic signature across the different subsites ($p<0.0001$), with significantly higher values for oral cavity compared to oropharynx ($p<0.0001$ for both HPV status) and larynx ($p<0.0001$). Figure 4d shows a statistically significant difference in the signature distribution between HPV-positive and negative cases ($p<0.0001$). The equivalent analysis done on the individual features is reported in the supplementary material A (Figures S13-S17).

In general, the addition of the radiomic signature to the clinical model improved the performances for both HPV+ and HPV- tumors. In HPV+ tumors, the C-index for OS raised from 0.75 [95% CI: 0.59-0.87] to 0.80 [95% CI: 0.59-0.87] ($C_{i,comb} > C_{i,clin}$ in 73 iterations, t = 0.74) and for DFS it raised from 0.64 [95% CI: 0.51-0.76] to 0.65 [95% CI: 0.51-0.78] ($C_{i,comb} > C_{i,clin}$ in 55 iterations, t = 0.08). In HPV- tumors, the C-index for OS raised from 0.63 [95% CI: 0.56-0.68] to 0.65 [95% CI: 0.59-0.72] ($C_{i,comb} > C_{i,clin}$ in 77 iterations, t = 0.84) and for DFS raised from 0.56 [95% CI: 0.53-0.59] to 0.61 [95% CI: 0.54-0.67] ($C_{i,comb} > C_{i,clin}$ in 87 iterations, t = 1.22). In all cases, the difference in survival curves of high/low risk groups is larger when using the combined model (Figure 4). The equivalent results for the training set are displayed in the supplementary material A (Figure S18). Calibration curves on the combined models on the test set are also displayed in the supplementary material A (Figure S19).

After performing the 2-way ANOVA, the only factor significantly affecting the signature value was tumor site ($p<0.0001$ and $p=0.0002$ in alive and dead patients, respectively). Vendor was not a significant factor. No significant interaction between factors (tumor site and scanner) was found.

Discussion

We developed an MRI-radiomics based prognostic signature for locally advanced HNC and successfully validated it on an external, prospective, multi-centric and multi-site validation dataset. We also demonstrated that the radiomic signature may be used as an independent prognostic factor, since the radiomics-based classification separates patients by OS and DFS for different sites, stages and HPV-status.

The radiomic signature was composed of features derived from both T1w and T2w images. We adopted these sequences since they are universally used in the MRI diagnosis of HNC. This strategy allowed for the size maximization of the study sets, and for the applicability of our models in other datasets. Our signature consists of 5 features. Among them, one is voxel volume (primary tumor),

which underlines the prognostic significance of tumor volume possibly in addition to TNM system. The others are related to the dispersion and non-uniformity of grey levels in the images. The feature *T1w-waveletLHL-firstorder-90Percentile* is related to the intensity of the high-pass filtered signal in T1w, while the features on the T2w account for the differences in texture (the grey level non-uniformity normalized) or for ranges of intensities (range and inter-quartile range). The only textural-related feature appearing in the signature (*waveletHHL_glrlm_GrayLevelNonUniformityNormalized*) is extracted from a T2w sequence and it is positively correlated with the heterogeneity of the ROI Moreover, an increased *WaveletHHL_glrlm_GrayLevelNonUniformityNormalized* is associated with an increase in the radiomic signature value, which translates into a worse OS. Studies have hypothesized that tumor characteristics at the cellular and genetic levels are reflected in the phenotypic patterns that can be captured by medical images (24). Indeed our radiomic seems to reflect tumor heterogeneity, a complex phenomenon that may impact on treatment outcome (25). Studies on MRI radiomics have already been reported. (8) (9). Mes et al. (8) have used a mono-centric dataset in which T1w images without contrast were acquired using different scanners, vendors and image acquisition parameters. They managed to train four different prognostic models for two different outcomes (OS and DFS) and two head and neck subsites (oral cavity and HPV- oropharynx). The models were validated on external mono-centric dataset giving an integrated AUC in the range 0.69-0.74. Bos et al. (9) used a mono-centric, multi-scanner and multi-protocol dataset of oropharyngeal cancer patients (T1w pre-and post-contrast and T2w) to develop a logistic regression classifier to predict OS and 2-years loco-regional recurrence (LRC). The models were evaluated through cross-validation and performed better in HPV+ (AUC_{LRC}=0.74 and AUC_{OS}=0.65) than in HPV- (AUC_{LRC}=0.65 and AUC_{OS}=0.59). A general limitation of MRI-radiomic studies is that they often focus on one or few specific anatomic sites, while our study included tumors from four head and neck subsites. This approach may lend to a wider usability of the identified signature in the field of HNC.

Moreover, our validation data were prospectively collected, which is a strong point in favor of this study. As a matter of fact, only a minority of the radiomic studies has an external validation set, and even less a prospective one.

In our study we classified tumors with the 8th edition of the TNM staging system and we focused on locally advanced tumors. Therefore, we set the premises for testing the proposed signature against the most valuable comparators in terms of prognostic forecasting.

The signature score distribution differed according to stage, subsite, and HPV status, consistently with available literature (17, 26).

Interestingly similar findings has been described with CT and PET radiomic studies as well (7, 27, 28). Our signature proved to be an independent prognostic factor for OS, but it forecasted DFS as well, even though with a lower performance. In general, stratification was less robust in groups with small sample size (larynx and hypopharynx) or low number of events, especially in HPV+ oropharynx. This limitation may be overcome in the future using larger validation subsite-specific sets with longer follow-up.

The features in this manuscript were extracted using Pyradiomics which has deviation from the IBSI recommendation. The known discrepancies between Pyradiomics and the IBSI guidelines are related to binning and resampling techniques. For as far as the binning is concerned, the issues are related to fix bin width discretization, which we did not used in this study. For as far as the resampling is concerned, there may be issues concerning grid alignment, value rounding and interpolator. All these factors may indeed cause modifications in the ROIs, but those should be mitigated by the fact that we are only using features that are stable to geometrical modifications.

We built our models based on features extracted from the primary tumor, not considering radiomic features from metastatic regional lymph nodes. We may not exclude that lymph node MRI-radiomics could provide additional prognostic information (29). Also, MRI-radiomics analyzed on healthy

tissues (e.g., neck muscles for the assessment of sarcopenia (30, 31)) may add further knowledge as well.

A potential point to address in our study is the use of missing data imputation in the data of the training set which did not have T1w or T2w images. Such missing data imputation may have confounding effects on the results. However, being the data generated randomly. it is unlikely that the confounding effect is a positive bias that creates optimistic results.

Another potential limitation to be addressed is the fact that we used more data coming from the Italian clinical centers compared to German and Dutch ones (See supplementary material A, "consort diagrams" and "MRI image acquisition parameters"). As a matter of fact, Italian centers provide more MRI, since that imaging technique is used for diagnosis and staging, while this is not the case for Germany and the Netherlands. This is a potential source of bias, but there is no dependency between the value of the radiomic signature and the clinical center, as proven by the analysis of variance as described in the results section.

This study may provide one of the first milestones in the path to a widespread use of MRI-radiomics in HNC.

References

1. Ferlay J, Colombet M, Soerjomataram I, et al.: Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur J Cancer* 2018; 103:356–387.

2. Zhang H, Graham CM, Elci O, et al.: Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy. *Radiology* 2013; 269:801–809.

3. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014; 5:4006.

4. Leijenaar RTH, Carvalho S, Hoebers FJP, et al.: External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol (Madr)* 2015; 54:1423–1429.

5. Leger S, Zwanenburg A, Pilz K, et al.: A comparative study of machine learning methods for time-To-event survival data for radiomics risk modelling. *Sci Rep* 2017; 7:1–11.

6. Bogowicz M, Riesterer O, Stark LS, et al.: Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. *Acta Oncol (Madr)* 2017; 56:1531–1536.

7. Zhai T, Dijk LV Van, Huang B, 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; 124:256–262.

8. Mes SW, van Velden FHP, Peltenburg B, et al.: Outcome prediction of head and neck squamous cell carcinoma by MRI radiomic signatures. *Eur Radiol* 2020; 30:6311–6321.

9. Bos P, van den Brekel MWM, Gouw ZAR, et al.: Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models. *Eur J Radiol* 2021; 139:109701.

10. Wong AJ, Kanwar A, Mohamed AS, Fuller CD: Radiomics in head and neck cancer: From exploration to application. *Transl Cancer Res* 2016; 5:371–382.

11. Liu Z, Wang S, Dong D, et al.: The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges. *Theranostics* 2019; 9:1303–1322.

12. Cavalieri S, De Cecco L, Brakenhoff RH, et al.: Development of a multiomics database for personalized prognostic forecasting in head and neck cancer: The Big Data to Decide EU Project. *Head Neck* 2020(October):1–12.

13. Amin MB, Edge SB, Greene FL, et al.: *AJCC Cancer Staging Manual*. Springer International Publishing: American Joint Commission on Cancer; 2017.

14. Jung F, Steger S, Knapp O, Noll M, Wesarg S: COSMO-coupled shape model for radiation therapy planning of head and neck cancer. In *Work Clin Image-Based Proced Transl Res Med Imaging CLIP 2014*; 2014:25–32.

15. Bologna M, Corino V, Mainardi L: Technical Note : Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain. *Med Phys* 2019; 46:5116–5123.

16. Tustison NJ, Cook PA, Gee JC: N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 2010; 29:1310–1320.

17. Leijenaar RT, Bogowicz M, Jochems A, et al.: Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. *Br J Radiol* 2018; 91:20170498.

18. van Griethuysen JJM, Fedorov A, Parmar C, et al.: Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* 2017; 77:e104–e107.

19. Pyradiomics features description [https://pyradiomics.readthedocs.io/en/2.1.0/features.html]

20. Bologna M, Corino VDA, Montin E, et al.: Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images. *J Digit Imaging* 2018; 31:879–894.

21. Harrell FE, Kerry LL, Mark DB: Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15:361–387.

22. Peto R, Peto J: Asymptotically Efficient Rank Invariant Test Procedures. *J R Stat Soc* 1972; 135:185–207.

23. Kaplan EL, Meier P: Nonparametric estimation from incomplete samples. *J Am Stat Assoc* 1958; 73:457–481.

24. Yip SSF, Aerts HJWL: Applications and limitations of radiomics. *Phys Med Biol* 2016; 61:R150–R166.

25. Dagogo-Jack I, Shaw AT: Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 2018; 15:81–94.

26. Liang C, Huan Y, He L, et al.: The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. *Oncotarget* 2016; 7:31401–31412.

27. Feliciani G, Fioroni F, Grassi E, et al.: Radiomic profiling of head and neck cancer: 18F-FDG PET texture analysis as predictor of patient survival. *Contrast Media Mol Imaging* 2018; 2018:1–9.

28. Bologna M, Corino V, Calareso G, et al.: Baseline mri-radiomics can predict overall survival in non-endemic ebv-related nasopharyngeal carcinoma patients. *Cancers (Basel)* 2020; 12:1–20.

29. Bogowicz M, Tanadini-Lang S, Guckenberger M, Riesterer O: Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. *Sci Rep* 2019; 9:1–7.

30. Wong A, Zhu D, Kraus D, Tham T: Radiologically Defined Sarcopenia Affects Survival in Head and Neck Cancer: A Meta-Analysis. *Laryngoscope* 2021; 131:333–341.

31. Dong X, Dan X, Yawen A, et al.: Identifying sarcopenia in advanced non-small cell lung cancer patients using skeletal muscle CT radiomics and machine learning. *Thorac Cancer* 2020; 11:2650–2659.

Accepted manuscript at https://doi.org/10.1016/j.radonc.2023.109638

Table 1 *Clinical data of the patients in the training and validation set. A column is also available displaying the p-value of a proper statistical test (Chi-squared, Mann-Whitney, Fisher's exact test) to identify significantly different distributions in the training and validation set.*

Patient characteristics	**Training set (285)**	**Validation set (234)**	**p-value**
Date of diagnosis	2008-2014	2015-2017	-
Median follow-up	63.38 m (95% CI 60.06-65.42)	27.66 m (95% CI 26.48-28.32)	<0.0001
Gender M F	 196 (69%) 89 (31%)	 161 (69%) 73 (31%)	1.0
Median age (range)	60 y (20-93)	61 y (20-93)	0.428
Primary tumor site Oral cavity Oropharynx HPV- Oropharynx HPV+ Hypopharynx Larynx	 133 (47%) 37 (13%) 80 (28%) 15 (5%) 20 (7%)	 108 (46%) 24 (10%) 68 (29%) 9 (4%) 25 (11%)	0.491
AJCC 7th edition III IV	 55 (19%) 230 (81%)	 34 (15%) 200 (85%)	0.161
AJCC 8th edition HPV- HNSCC III IV p16+ Oropharynx I II III	 53 (26%) 152 (74%) 30 (38%) 21 (26%) 29 (36%)	 30 (18%) 136 (82%) 30 (44%) 15 (22%) 23 (34%)	 0.08 0.696
Smoking status Current/former Never Unknown	 208 (73%) 77 (27%) -	 136 (58%) 65 (28%) 32 (14%)	0.224
ACE27 0 1 2 3 Unknown	 87 (31%) 131 (46%) 57 (20%) 10 (3%) -	 60 (26%) 101 (43%) 29 (13%) 10 (4%) 33 (14%)	0.368
Treatments* Surgery Radiotherapy Chemotherapy	 154 (54%) 250 (88%) 186 (65%)	 120 (56%) 174 (88%)** 124 (58%)	 0.785 1.0 0.077

Outcome events OS DFS	 112 (39%) 140 (49%)	 46 (20%) 79 (34%)	 <0.0001 0.0005

Legend: 95% CI, 95% confidence interval; HNSCC, head and neck squamous cell carcinoma; HPV, human papilloma virus; m, months; OPC, oropharyngeal carcinoma; y, years

* details about treatments were available for 284 cases of the training set and for 215 of the validation set

**data about radiotherapy was available in 198 prospective cases

Table 2 *Regression coefficients, hazard ratios and p-values for the features used for the radiomic, clinical (only HPV-) and combined models. For tumor subsite, Hypopharynx is used as the reference category.*

Model	Feature	Beta Coefficient	Hazard ratio	p-value
Radiomic	RF1	-0.25 [-0.48 /-0.02]	0.78 [0.62/ 0.98]	0.03
	RF2	0.14 [-0.06 / 0.34]	1.15 [0.94/ 1.40]	0.21
	RF3	0.08 [-0.15 / 0.31]	1.08 [0.86/ 1.36]	0.49
	RF4	0.13 [-0.12 / 0.38]	1.14 [0.89/ 1.46]	0.30
	RF5	0.20 [-0.12/ 0.36]	1.22 [0.89/ 1.43]	0.10
Clinical (HPV+)	Stage TNM 8th	0.44 [-0.30/ 1.18]	1.55 [0.74/ 3.25]	0.24
Clinical (HPV-)	Stage TNM 8th	0.80 [0.27/ 1.33]	2.23 [1.31/ 3.78]	0.003
	Is_Larynx	-0.90 [-2.08/ 0.18]	0.41 [0.12/ 1.20]	0.13
	Is_OralCavity	-0.04 [-0.67/ 0.67]	0.96 [0.51/ 1.95]	0.99
	Is_Oropharynx	-0.40 [-1.20/ 0.40]	0.67 [0.30/ 1.49]	0.32
Combined (HPV+)	Radiomic signature	1.91 [0.09/ 3.73]	6.75 [1.09/ 41.68]	0.04
	Stage TNM 8th	0.37 [-0.37/ 1.11]	1.45 [0.69/ 3.03]	0.33
Combined (HPV-)	Radiomic signature	0.67 [0.28/ 1.06]	1.95 [1.32/ 2.89]	<0.001
	Stage TNM 8th	0.73 [0.18/ 1.28]	2.08 [1.20/ 3.60]	0.009

	Is_Larynx	-0.84 [-2.02/ 0.34]	0.43 [0.13/ 1.40]	0.16
	Is_OralCavity	-0.27 [-0.96/ 0.42]	0.76 [0.38/ 1.52]	0.44
	Is_Oropharynx	-0.38 [-1.18/ 0.42]	0.68 [0.31/ 1.52]	0.35

RF1: *T1w_waveletLHL_firstorder_90Percentile*
RF2: T2w_original_shape_VoxelVolume
RF3: T2w_waveletHHL_glrlm_GrayLevelNonUniformityNormalized
RF4: T2w_waveletLLL_firstorder_InterquartileRange
RF5: T2w_waveletLLL_firstorder_Range

Figure 1 *Consort diagram for the patients' selection in the training retrospective cohort (a) and in the validation prospective cohort (b).*

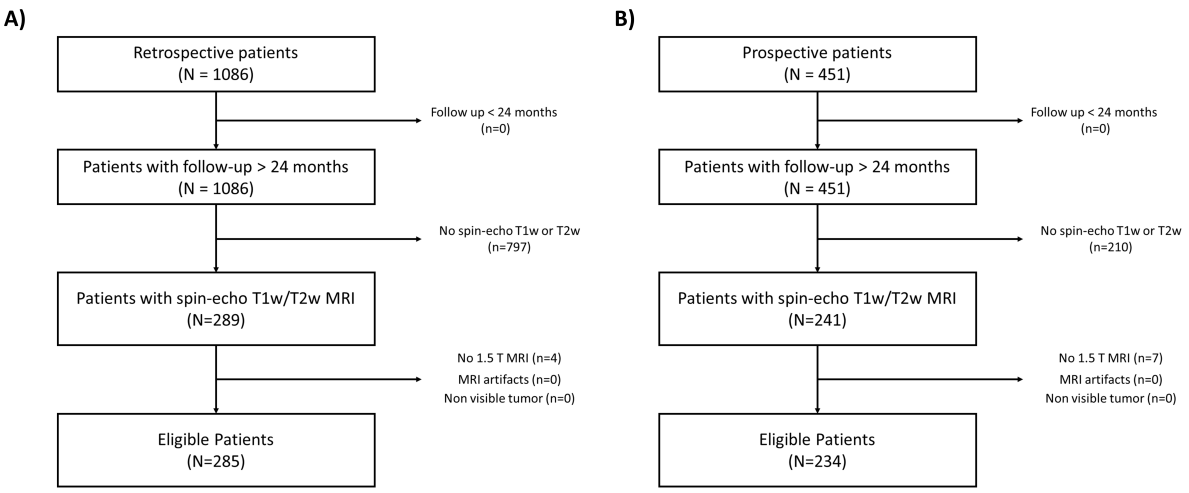
Figure 2 *Kaplan-Meier curves for all the patients of the validation set. a) Overall survival (OS) by volume-based risk groups. b) Disease-free survival (DFS) by volume-based risk groups. c) OS by radiomics-based risk groups. d) DFS by radiomics-based risk groups. The p-values of log-rank test comparing the survival curves are also displayed. The lower portion of each plot represents the number of patients remained in the study along time.*

Figure 3 *Radiomic signature grouped by different prognostic factors. a) Stage TNM 8th (for HPV+ cancers). b) Stage TNM 8th edition (for HPV- cancers). c) HPV status. d) Tumor site. The boxplots refer to merged training and validation sets. Significant differences are marked by asterisk.*

Figure 4 *Kaplan-Meier curves for the patients of the validation set, divided in high and low risk groups (dashed and continuous lines respectively), using a signature obtained from clinical variables alone (red) and a signature combining both clinical and radiomic features (blue). The p-values for the log-rank test are also displayed using the corresponding colors. a-b) Overall survival (OS) and disease-free survival (DFS) for HPV+ tumors. c-d) OS and DFS for HPV- tumors.*

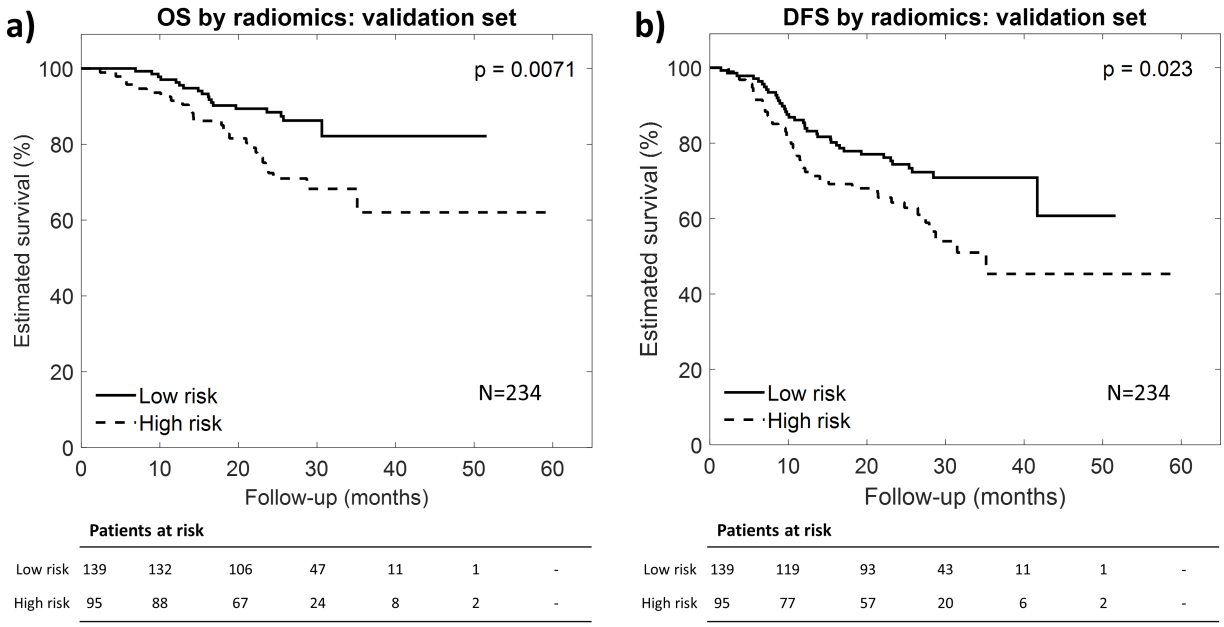
Accepted manuscript at https://doi.org/10.1016/j.radonc.2023.109638

Figure 1



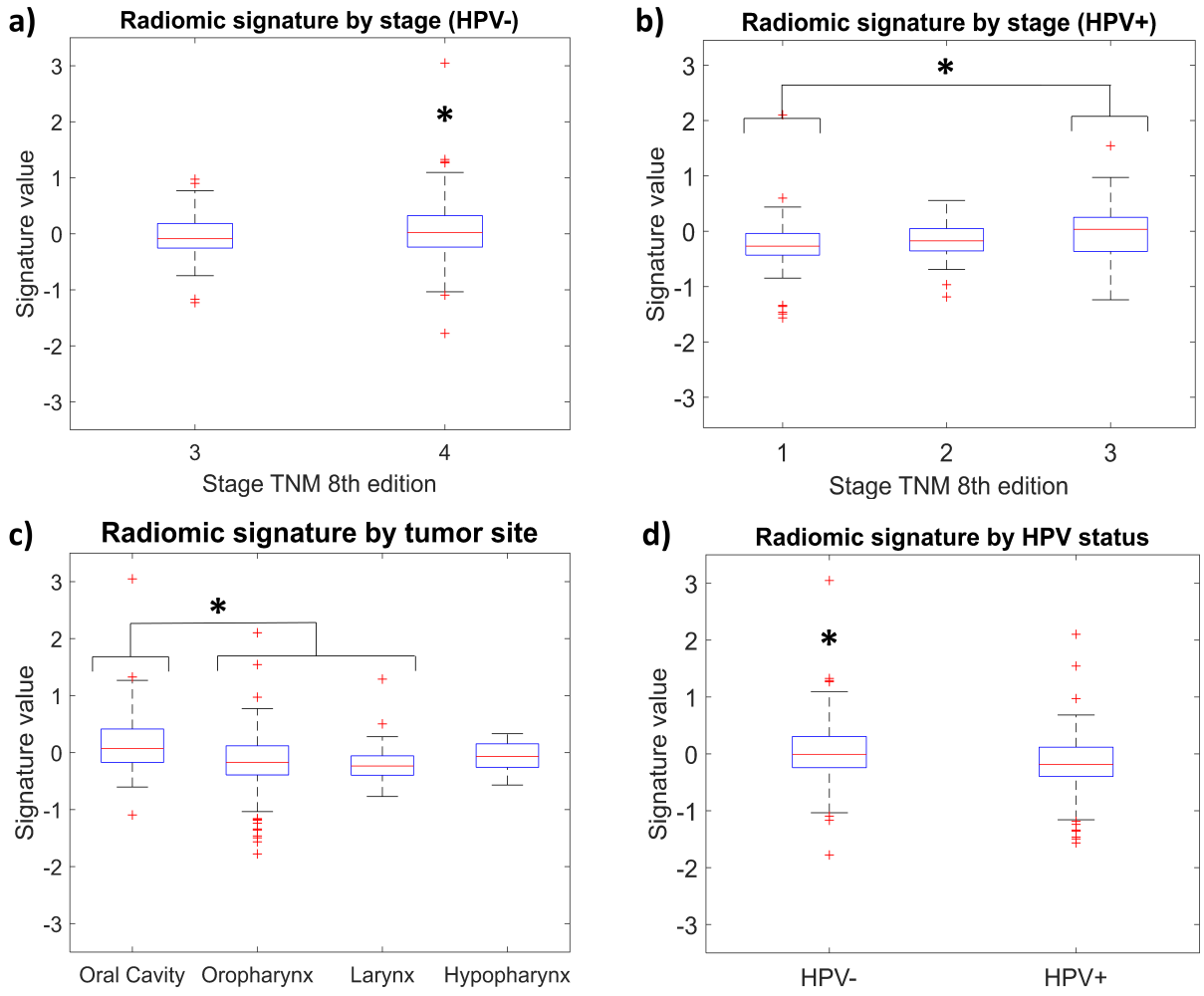
Accepted manuscript at https://doi.org/10.1016/j.radonc.2023.109638

Figure 2



Accepted manuscript at https://doi.org/10.1016/j.radonc.2023.109638

Figure 3



Accepted manuscript at https://doi.org/10.1016/j.radonc.2023.109638

Figure 4