Clinical and Translational Radiation Oncology 26 (2021) 62–70

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

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

Definition and validation of a radiomics signature for loco-regional tumour control in patients with locally advanced head and neck squamous cell carcinoma

Asier Rabasco Meneghetti [a,b], Alex Zwanenburg [b,c,d], Stefan Leger [b,c,d], Karoline Leger [b,c,d,e], Esther G.C. Troost [a,b,c,d,e], Annett Linge [b,c,d,e], Fabian Lohaus [b,c,d,e], Andreas Schreiber [f], Goda Kalinauskaite [g,h], Inge Tinhofer [g,h], Nika Guberina [i,j], Maja Guberina [i,j], Panagiotis Balermpas [k,l], Jens von der Grün [k,l], Ute Ganswindt [m,n,o,p], Claus Belka [m,n,o], Jan C. Peeken [m,q,r], Stephanie E. Combs [m,q,r], Simon Böke [s,t], Daniel Zips [u], Mechthild Krause [a,b,c,d,e], Michael Baumann [a,b,c,d,e,u], Steffen Löck [b,c,e,*]

[a] Institute of Radiooncology – OncoRay, Helmholtz-Zentrum Dresden – Rossendorf, Dresden, Germany
[b] OncoRay - National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany
[c] German Cancer Consortium (DKTK), Partner Site, Dresden, Germany and German Cancer Research Center (DKFZ), Heidelberg, Germany
[d] National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, and : Helmholtz Association/Helmholtz-Zentrum Dresden – Rossendorf (HZDR), Dresden, Germany
[e] Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany
[f] Department of Radiotherapy, Hospital Dresden-Friedrichstadt, Dresden, Germany
[g] German Cancer Consortium (DKTK) Partner Site Berlin, Germany and German Cancer Research Center (DKFZ), Heidelberg, Germany
[h] Department of Radiooncology and Radiotherapy, Charité University Medicine Berlin, Germany
[i] German Cancer Consortium (DKTK), Partner Site Essen, Germany and German Cancer Research Center (DKFZ), Heidelberg, Germany
[j] Department of Radiotherapy, University Hospital Essen, Medical Faculty, University of Duisburg-Essen, Essen, Germany
[k] German Cancer Consortium (DKTK), Partner Site Frankfurt, Germany and German Cancer Research Center (DKFZ), Heidelberg, Germany
[l] Department of Radiotherapy and Oncology, Goethe-University Frankfurt, Germany
[m] German Cancer Consortium (DKTK), Partner Site Munich Germany, German Cancer Research Center (DKFZ), Heidelberg, Germany
[n] Department of Radiation Oncology, Ludwig-Maximilians-Universität, Munich, Germany
[o] Clinical Cooperation Group, Personalized Radiotherapy in Head and Neck Cancer, Helmholtz Zentrum, Munich, Germany
[p] Department of Radiation Oncology, Medical University of Innsbruck, Anichstraße 35, A-6020 Innsbruck, Austria
[q] Department of Radiation Oncology, Technische Universität München, Germany
[r] Institute of Radiation Medicine (IRM), Helmholtz Zentrum München, Neuherberg, Germany
[s] German Cancer Consortium (DKTK), Partner Site Tübingen, Germany German Cancer Research Center (DKFZ), Heidelberg, Germany
[t] Department of Radiation Oncology, Faculty of Medicine and University Hospital Tübingen, Eberhard Karls Universität Tübingen, Germany
[u] German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

Article history:
Received 18 September 2020
Revised 18 November 2020
Accepted 21 November 2020
Available online 27 November 2020

Keywords:
HNSCC
Radiomics
Validation
Biomarker
Machine learning
Loco-regional control

ABSTRACT

Purpose: To develop and validate a CT-based radiomics signature for the prognosis of loco-regional tumour control (LRC) in patients with locally advanced head and neck squamous cell carcinoma (HNSCC) treated by primary radiochemotherapy (RCTx) based on retrospective data from 6 partner sites of the German Cancer Consortium - Radiation Oncology Group (DKTK-ROG).
Material and methods: Pre-treatment CT images of 318 patients with locally advanced HNSCC were collected. Four-hundred forty-six features were extracted from each primary tumour volume and then filtered through stability analysis and clustering. First, a baseline signature was developed from demographic and tumour-associated clinical parameters. This signature was then supplemented by CT imaging features. A final signature was derived using repeated 3-fold cross-validation on the discovery cohort. Performance in external validation was assessed by the concordance index (C-Index). Furthermore, calibration and patient stratification in groups with low and high risk for loco-regional recurrence were analysed.
Results: For the clinical baseline signature, only the primary tumour volume was selected. The final signature combined the tumour volume with two independent radiomics features. It achieved moderately

* Corresponding author at: OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany.
E-mail address: Steffen.Loeck@OncoRay.de (S. Löck).

https://doi.org/10.1016/j.ctro.2020.11.011

good discriminatory performance (C-Index [95% confidence interval]: 0.66 [0.55–0.75]) on the validation cohort along with significant patient stratification (p = 0.005) and good calibration.
Conclusion: We identified and validated a clinical-radiomics signature for LRC of locally advanced HNSCC using a multi-centric retrospective dataset. Prospective validation will be performed on the primary cohort of the HNprädBio trial of the DKTK-ROG once follow-up is completed.

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) is the fifth most incident tumour entity worldwide. Patients suffering from locally advanced stages show a 5-year survival rate of approximately 50% [1]. In order to stratify patients for different treatment options, biomarkers reflecting individual tumour aggressiveness and response to treatment are required [2]. For HNSCC, several validated biomarkers for the prognosis of treatment outcome have been developed including tumour volume [3–6], human papilloma virus (HPV) status [7,8] and gene signatures, e.g. for hypoxia [6,9–12]. Biomarkers reflecting tumour heterogeneity may help to further improve patient stratification.

Radiomics performs a quantitative characterisation of medical imaging to identify image biomarkers. It employs machine learning algorithms for the evaluation of disease diagnosis or for the prognosis of treatment outcome and has been applied to several tumour entities and different imaging modalities [13]. In HNSCC, radiomics has been applied, e.g. to assess local tumour control using pre-treatment positron-emission tomography (PET) and computerised tomography (CT) [14,15] or for HPV status prediction with PET and CT imaging [16,17]. A radiomics signature for overall survival of HNSCC has been developed using CT imaging, and was externally validated [18,19]. Radiomics has also been used to analyse in-treatment CT images for the prognosis of loco-regional tumour control (LRC) [20], enhancing pre-treatment-only models.

In our previous work by Leger et al. [21], we aimed to compare different machine learning algorithms and feature selection methods in a radiomics analysis for the endpoint LRC in locally advanced HNSCC based on pre-treatment CT data. We identified a subset of algorithms that may be applied for radiomics studies to capture the observed variability in the data. In the present study, we used these algorithms to develop a specific signature containing clinical parameters and radiomics features for the prognosis of LRC in locally advanced HNSCC after primary radiochemotherapy (RCTx). We applied a modified version of the workflow presented in [21] consisting of stability analysis, feature clustering, feature selection, hyperparameter optimisation, model building and independent validation.

2. Material and methods

2.1. Patient cohort

Radiomics signatures were developed and validated based on 318 patients. All patients were diagnosed with advanced HNSCC, confirmed by histopathology, and underwent primary RCTx with curative intent. Dose was prescribed to the tumour region and adjacent lymph nodes. Treatment doses of up to 76.8 Gy were delivered in different hyperfractionated, accelerated schedules (66% of patients) or up to 77 Gy in normofractionated schedules (34%) with different boost concepts. Concomitant cisplatin (96%) or mitomycin C (4%) was applied in combination with 5-Fluorouracil. Patients were allocated to a discovery cohort (n = 233) and to a validation cohort (n = 85) based on the different included studies rather than on treatment centre, similar to [21] (Supplementary Table 1). 147 of the discovery patients were treated in one of six partner sites of the German Cancer Consortium - Radiation Oncology Group (DKTK-ROG) between 2005 and 2011 [6]. 86 patients were treated at the University Hospital (UKD, Dresden) between 2002 and 2014 [3]. 51 patients in the validation cohort were treated within a prospective trial (NCT00180180) at the UKD between 2006 and 2013 [12,22], 20 were treated at the UKD and the Radiotherapy Centre Dresden-Friedrichstadt between 2005 and 2009, the remaining 14 were treated at the Department of Radiation Oncology of the University Hospital Tübingen, between 2008 and 2013 [23].

All analyses were carried out in accordance with the relevant ethical and legal guidelines and regulations. Ethical approval for the multicentre retrospective analyses of clinical and imaging data was obtained from the Ethics Committee of the Technische Universität Dresden, Germany, EK177042017.

2.2. Study design

Fig. 1 presents the design of our study. The primary endpoint was LRC, which was calculated from the first day of RCTx to the day of event or censoring. First, we identified a clinical signature prognostic for LRC that was based on clinical parameters only, using the discovery cohort. Then we used radiomics features com-

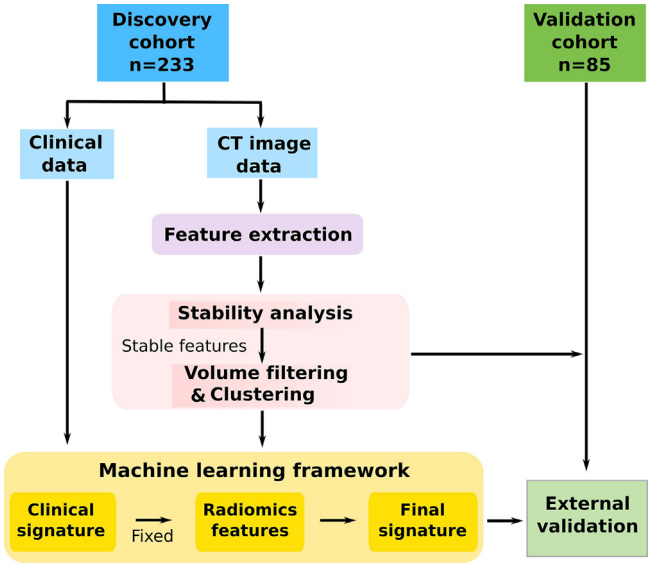

Fig. 1. Representation of the study design. First, a clinical baseline signature prognostic for loco-regional tumour control was developed using the discovery cohort within the machine learning framework. This signature was then supplemented by CT radiomics features from the discovery cohort to develop a final clinical-radiomics signature. Radiomics features were subjected to stability analysis, volume filtering and clustering before entering the machine learning framework for prognostic modelling. The final signature was externally validated on the validation cohort.

puted from the primary gross tumour volume (GTV) delineated in pre-treatment CT imaging of patients in the discovery cohort to supplement the clinical signature and create a final clinical-radiomics signature. A prognostic model was trained in the discovery cohort for both signatures. These models were then assessed on the validation cohort by calculating the concordance index (C-Index) as a prognostic measure, by analysing model calibration and by stratifying patients into groups of low and high risk for loco-regional recurrence.

2.3. Image pre-processing and feature extraction

Patients received a CT scan for treatment planning prior to radiotherapy. Image acquisition and reconstruction parameters are summarised in Supplementary Table 2. The GTV was delineated in each scan by experienced radiation oncologists at our institution. Voxels in each CT volume were resampled to an isotropic size of $1.0 \times 1.0 \times 1.0$ mm^3 using cubic splines to compensate for differing voxel spacing and slice thickness between centres. Spatial filtering techniques were applied to the base image after resampling to quantify characteristics such as edges or blobs. A set of Laplacian of Gaussian (LoG) filters with 5 different kernel widths (1 mm, 2 mm, 3 mm, 4 mm, 5 mm) were applied individually to the base image. The five response maps were averaged to a single image. The entire image pre-processing pipeline was implemented according to the recommendations by the Image Biomarker Standardisation Initiative (IBSI) [24,25].

From the base image and the LoG-transformed image, a set of 18 statistical, 2 local-intensity based, 29 morphological, 37 intensity-histogram-based and 137 texture-based features were extracted from the GTV leading to 446 features per patient. All features were calculated using a 3D approach. Features were computed in full compliance with the IBSI [25]. Radiomics image processing and feature computation parameters are summarised in Supplementary Table 3. The entire image processing and extraction process was done using the publicly available MIRP Python package [26].

2.4. Stability and clustering of features

Radiomics features should be robust to different sources of variation, e.g. acquisition parameters and positioning uncertainties, to allow for external reproduction. As proposed previously, image augmentation techniques can be used to identify non-robust features [27]. For the present study, GTVs were rotated (-4°, −2°, 0°, 2°, 4°) and volume-changed (-20%, −10%, 10%, 20%) in the discovery cohort, producing 20 new images per patient from which to analyse individual feature stability. The intra-class correlation coefficient (ICC) was calculated with 95% confidence interval (CI), quantifying the similarity of feature values under different perturbations for every feature [28]. For dimensionality reduction and to only use robust features for model-building, features with the lower boundary of the 95% CI of the ICC below 0.75 were excluded. The same features were excluded in the validation cohort.

We subsequently computed the Spearman correlation coefficient ρ between robust radiomics features and features in the clinical signature. Every radiomics feature with $|\rho| \geq 0.6$ to one of the features in the clinical signature was removed. Then, we identified radiomics features that were highly similar by assessing their mutual correlation. Features were clustered together using hierarchical clustering with complete linkage and $1 - |\rho|$ as the distance metric. All features with $|\rho| \geq 0.8$ were clustered by cutting the hierarchical tree at height 0.2. The feature of each cluster with maximum mutual information with the outcome was chosen as the representative based on the discovery cohort. The same features were selected for the validation cohort.

2.5. Identifying a clinical and clinical-radiomics signature

Two signatures were developed. First, we created a prognostic clinical signature based on clinical parameters only. Subsequently, we complemented this signature with radiomics features to create the final clinical-radiomics signature. Both signatures were developed using an in-house end-to-end statistical learning software package. The 4 major processing steps of this package are shown in Fig. 2: (i) feature pre-processing, (ii) feature selection (detailed workflow in Supplementary Fig. 1), (iii) hyper-parameter optimisation for the machine learning algorithms, and (iv) model building with internal validation.

Overall, steps (i)-(iii) were performed within 33 repetitions of 3-fold cross validation [29] nested in the discovery cohort to identify an optimal signature, i.e. the steps were repeatedly performed using the training runs and validated on the validation runs of the cross-validation runs of the discovery cohort.

(i) Features were transformed using the Yeo-Johnson transformation to align their distribution to a normal distribution [30]. Afterwards, features were z-transformed to mean zero and standard deviation one. Both transformations were performed on the training runs and the resulting transformation parameters were applied unchanged to the features in the validation fold.

(ii) Based on the results from [21], three supervised feature-selection algorithms were considered: Spearman correlation (Spearman), minimal redundancy maximum relevance (MRMR) [31] and regularised Cox regression (Lasso) [32]. Subsequently, the features selected by each of these methods were used by three different prognostic models: Cox regression (Cox), boosted Cox regression (BGLM-Cox) [33], and random survival forests (RSF) [34]. All models can work with continuous time-to-event data.

(iii) In order to reduce overfitting in our models, hyperparameters were tuned automatically using the SMBO algorithm based on bootstrap sampling of the training runs for each model [35].

(iv) A signature was identified as follows. First, features were ranked according to their occurrence across the 99 cross-validation runs [36]. Occurrence was defined as the percentage of runs where a feature was found among the five most important features. The signature size was defined as the median signature size across the cross-validation runs. This procedure was conducted for every combination of feature selection method and model. The resulting nine signatures were then used to train prognostic models on 200 bootstraps of the entire discovery cohort in order to evaluate their discriminatory power based on the C-Index. The final signature was chosen based on the highest median C-Index on the out-of-bag (OOB) data.

The outlined procedure was first performed to identify a clinical baseline signature. Afterwards, a clinical-radiomics signature was developed by applying the same modelling procedure to a setup in which the clinical baseline signature was fixed and supplemented by CT imaging features. Both signatures were then used to train a model based on the entire discovery cohort by repeating steps (i) and (iii) with their respective machine learning algorithm, leading to a clinical model and a clinical-radiomics model. These models were subsequently assessed in the validation cohort.

2.6. Statistical analyses

LRC was compared between discovery and validation cohort via the log-rank test. Categorical variables of the clinical data were compared between discovery and validation cohorts by the χ^2 test, whereas continuous variables were compared using the Mann-Whitney-*U* test. All tests were conducted two-sided, except for the one-sided permutation test, at $p = 0.05$ level of significance on R software version 3.6.0 (R Core Team, 2019).

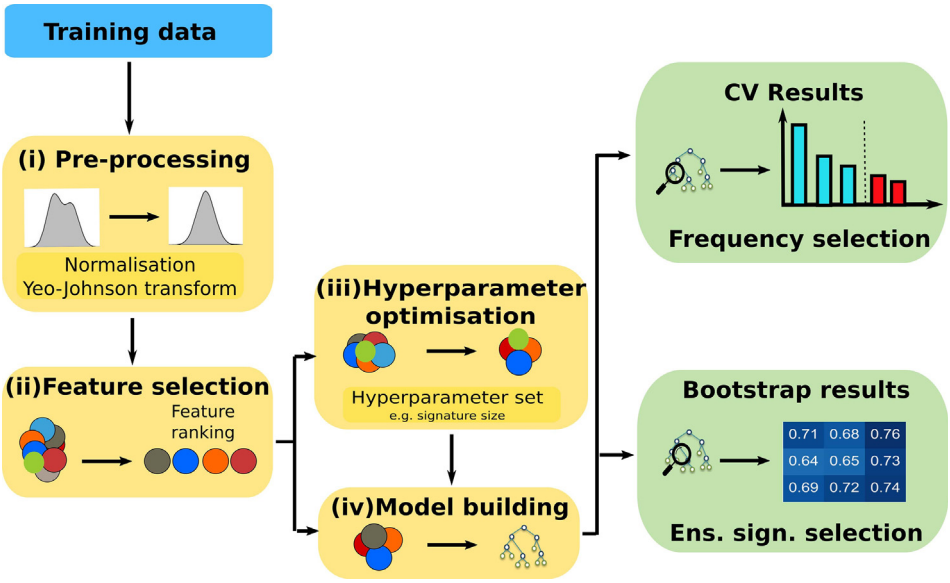

Fig. 2. Overview of the machine learning framework for signature selection. The major steps of the framework are: (i) pre-processing of the training dataset (the discovery cohort), (ii) feature selection, (iii) hyperparameter optimization and (iv) model building with internal validation. An ensemble signature was created for every combination of feature selection and machine learning algorithm by assessing feature occurrence across cross-validation runs in a 33-times repeated 3-fold cross-validation setting (Frequency selection). Models were created in 200 bootstrap subsamples of the discovery cohort with each corresponding ensemble signature. The concordance index (C-index) on the out-of-bag data was subsequently computed for each model. The signature and machine learning algorithm with the highest median C-index was selected, out of nine possible options (Ens. sign. selection). This process was applied first for clinical features. Then we repeated the procedure for CT-based radiomics features, supplemented by the selected clinical signature, to create a clinical-radiomics signature. Models based on both signatures were then assessed on the basis of discrimination, stratification and calibration in an external validation cohort.

Available clinical features were: GTV, age, total dose, gender, tumour localisation, UICC stage (2010), T stage, N stage, grading, p16 status, HPV16 DNA status, alcohol and smoking status. Missing values were imputed by their median value for numerical variables and by their mode for categorical variables except for alcohol consumption, smoking status and p16 status. These three features were transformed into two binary features each, representing positivity and non-availability. The following categorical variables were binarised: cT stage (0 for cT < 4 and 1 for cT = 4), cN stage (0 for cN < 2 and 1 cN $\geq$ 2), Grading (0 for Grading $\leq$ 2 and 1 for Grading > 2) and UICC stage (0 for UICC < 4, 1 for UICC = 4) as in a previous study [6].

Associations between the final model prognosis and LRC were evaluated based on the C-Index, for which the median value and 95% confidence interval were reported [37]. Patients were stratified into a low and high risk group using the optimised risk prediction of the discovery cohort [38]. LRC of these groups was estimated by Kaplan Meier curves, comparisons between groups were assessed with the log-rank test. Calibration at 24 months was assessed via the Greenwood Nam d'Agostino test (GND test) [39]. Correlations between features were assessed by the Spearman correlation coefficient (ρ). Permutation tests were performed to analyse the importance of the features in the final signatures: for 1000 bootstraps, one selected feature was randomly permuted. The resulting C-Index distribution on the discovery and validation cohort was used to define a heuristic p-value as the percentage of permuted C-Indexes greater than the unpermuted result. This procedure was repeated for every feature in the final signature. Differences in discriminatory performance between clinical and clinical-radiomics models were evaluated using bootstraps. We created 1000 bootstraps of the validation set, and for each bootstrap computed the C-Index for both models. We then determined a p-value by assessing the fraction of bootstraps for which the clinical model had a higher C-Index than the clinical-radiomics model.

3. Results

Clinical characteristics of the discovery and validation cohort are shown in Table 1. Median follow-up time was 15.8 months for the discovery cohort and 19.6 months for the validation cohort. The primary endpoint LRC was not significantly different between cohorts (p = 0.35, Supplementary Fig. 2). Patients in the validation cohort presented larger GTV (p = 0.060), younger age (p = 0.015), and were treated with a marginally higher dose (p < 0.001). Associations between clinical variables and LRC are shown in Supplementary Table 4.

First, a clinical baseline signature prognostic for LRC was developed. All applied feature selection methods selected GTV as the most important prognostic variable (80% occurrence or more), with other features rarely chosen (<30% occurrence) (Supplementary Table 5). The final clinical model was a univariate Cox regression model containing the GTV. This model showed a median C-Index of 0.59 with a 95% CI of [0.53–0.65] on the entire discovery cohort and a C-Index of 0.61 [0.51–0.71] on the validation cohort. Using an optimised cut-off of 0.982 (29.296 cm^3), the risk groups were significantly different in discovery (p = 0.002) and borderline significant in validation (p = 0.052). The model was well calibrated (discovery: GND = 0.76, slope = 1.09 [0.35–1.83], offset = -0.06 [-0.49–0.38]; validation: GND = 0.80, slope = 1.03 [0.51–1.55], offset = 0.05 [-0.24–0.34]). Information about model and transformation parameters can be found in Supplementary Table 6.

Afterwards, the final clinical-radiomics signature based on the GTV and additional CT radiomics features was developed on the

Table 1
Characteristics of clinical features for discovery (left) and validation (right) cohort along with p-values for homogeneity tests between cohorts.

Variable	Discovery cohort Median (range)		Validation cohort Median (range)	p-value
GTV (cm^3)	29.1 (1.3–321.7), Missing: 0		40.5 (2.7–238.8) Missing: 0	0.061
Age (years)	58.3 (39.2–84.5), Missing: 0		54 (37.0–76.0) Missing: 19	**0.015**
Total Dose (Gy)	72 (67.8–76.8), Missing: 28		72 (69–77) Missing: 4	**<0.001**
		Number of 233 (%)	**Number of 85 (%)**	
Gender	0 (Male)	194 (83.3)	77 (90.6)	0.43
	1 (Female)	39 (15.2)	8 (9.4)	
Tumour site	Oropharynx	101 (43.3)	29 (34.1)	0.38
	Hypopharynx	65 (27.9)	28 (32.9)	
	Larynx	8 (3.4)	5 (5.9)	
	Oral cavity	59 (25.4)	23 (27.1)	
UICC stage (2010)	1	0 (0)	1 (1.2)	0.079
	2	1 (0.4)	2 (2.4)	
	3	17 (7.3)	9 (10.5)	
	4	212 (90.9)	73 (85.9)	
	Missing	3 (1.3)	0	
cT stage	1	2 (0.8)	2 (2.3)	0.15
	2	24 (10.3)	9 (10.6)	
	3	58 (24.9)	30 (35.3)	
	4	146 (62.7)	48 (56.4)	
	Missing	3 (1.3)	0	
cN stage	0	38 (16.3)	10 (11.8)	0.23
	1	9 (3.9)	8 (9.4)	
	2	171 (73.4)	64 (75.3)	
	3	15 (6.4)	3 (3.5)	
Grading	0	1 (0.4)	0	0.051
	1	6 (2.6)	0	
	2	135 (57.9)	43 (50.6)	
	3	62 (26.6)	35 (41.2)	
	Missing	29 (12.5)	7 (8.2)	
p16 status	0 (Negative)	142 (60.9)	57 (67.1)	0.45
	1 (Positive)	91 (39.1)	28 (32.9)	
HPV16 DNA	0 (Negative)	184 (78.9)	34 (40.0)	1.00
	1 (Positive)	22 (9.4)	4 (4.7)	
	Missing	27 (11,7)	47 (55.3)	
Alcohol	0 (No)	69 (29.6)	23 (27.1)	0.72
	1 (Regular)	103 (44.2)	25 (29.4)	
	Missing	61 (26.2)	37 (43.5)	
Smoking	0 (Negative)	46 (19.7)	13 (15.3)	0.97
	1 (Positive)	185 (79.4)	51 (60.0)	
	Missing	2 (0.9)	21 (24.7)	

GTV: gross tumour volume, UICC: *Union international contre le cancer*, HPV: human papillomavirus, DNA: deoxyribonucleic acid

discovery cohort. 349 out of 446 stable CT features remained after performing the stability analysis, i.e. eliminating features with a lower boundary of the 95% CI of the ICC below 0.75. One hundred and forty-three of these features that were highly correlated with GTV ($|\rho| \geq 0.6$) were excluded. Clustering of intercorrelated features ($|\rho| \geq 0.8$) further reduced the number of these features to 61 (see Supplementary Table 7). In combination with the fixed feature GTV, these CT radiomics features were used in a 3-fold cross validation setting with 33 repetitions (99 runs) to assess LRC. Resulting C-Indices of nested training and validation results are presented in Fig. 3a and 3b, respectively. On average the nested validation C-Index was 0.59 and there was little variability between the modelling algorithms. Hyperparameter information for the different combinations can be found in Supplementary Table 8.

The best performing signature on 200 bootstraps of the discovery cohort had a signature size 3 and was identified using MRMR feature selection and the Cox-regression model (C-Index: 0.64 [0.58–0.69], Fig. 3 c,d). The included features were the GTV and the CT radiomics features log_ngl_hdhge (texture, occurrence: 30.3%) and stat_p10 (statistical, occurrence 20.2%). The two radiomics features were weakly correlated among themselves ($\rho = 0.40$) and with the GTV ($\rho \leq 0.51$). The feature Log_ngl_hdhge (IBSI: 9QMG) represents big groups of nearby voxels with similarly high intensity within the GTV and is derived from the lower right quadrant of the neighbouring grey level dependence matrix (NGLDM) in the LoG image. Feature stat_p10 (IBSI: QG58) is related to the intensity in the entire GTV and describes the 10th percentile intensity of the base image. The features are presented for example patients in Fig. 4.

The clinical-radiomics signature was finally trained in a Cox model (C-Index: 0.63 [0.58–0.69]) on the entire discovery cohort and was successfully validated on the validation cohort (C-Index: 0.66 [0.55–0.75]) for the endpoint LRC. It showed improved discriminatory power compared to the clinical model with a trend to statistical significance ($p = 0.076$). Details about model coefficients and transformation parameters can be found in Table 2.

Based on the validated model, patients were stratified into groups at high and low risk of loco-regional recurrence using the optimised risk cutoff value 1.343 of the discovery data. This cutoff was applied to the validation cohort. Stratified risk groups significantly differed in LRC in discovery and validation ($p < 0.001$ and $p = 0.005$, respectively; Fig. 5 a,b). The model was well calibrated in discovery (GND = 1.00, slope = 1.06 [0.71–1.41], offset = -0.04 [-0.24–0.17]) and validation cohorts (GND = 0.55, slope = 0.93 [0.27–1.59], offset = 0.12 [-0.24–0.47]) (Fig. 5 c,d). The baseline survival curve can be found in Supplementary Fig. 3. Concerning feature importance, permutation tests revealed that the all selected features contributed significantly or showed a statistical trend for association with LRC in discovery, while the tumour volume and stat_p10 were significantly associated with LRC in validation, see Supplementary Table 9.

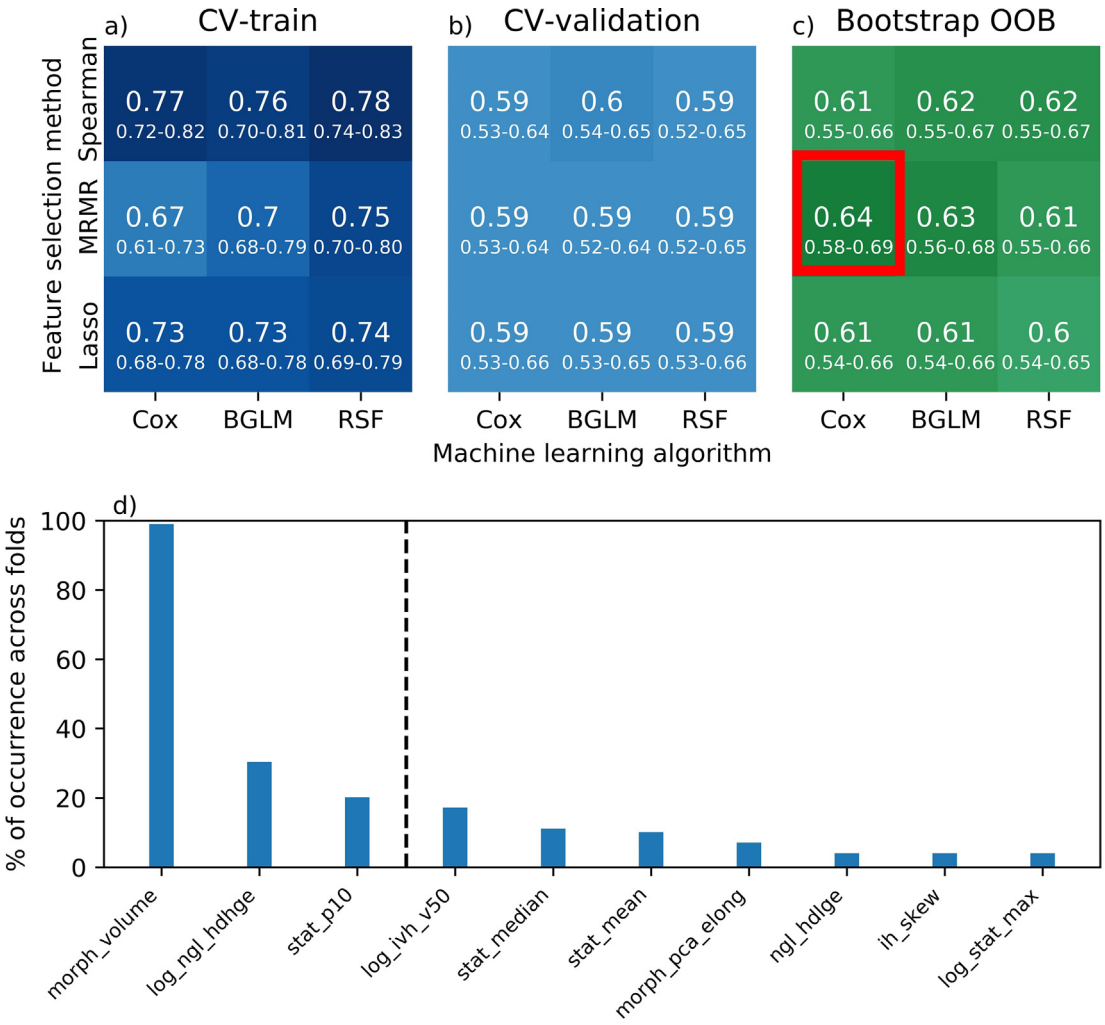

Fig. 3. C-Index of models based on different feature-selection methods and machine learning algorithms for the prognosis of loco-regional tumour control. Shown are the median (95% confidence interval) results from (a) the training runs and (b) the validation runs of 33 times repeated 3-fold cross validation (CV) as well as (c) the results of the ensemble signatures on 200 bootstrap samples of the discovery cohort evaluated on the out-of-bag data (OOB). In (d), the occurrences of radiomics features in CV are shown for the best performing signature (red box in (c)). The three features with the highest occurrence were selected for the final radiomics signature (dashed line). The primary tumour volume (morph_volume) was fixed and thus always occurred. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

In the present study, we developed a clinical-radiomics signature for the prognosis of LRC in patients with locally advanced HNSCC that received primary RCTx. Using stability analysis and clustering techniques combined with previously proposed machine learning algorithms and feature-selection methods, the final signature contained two features derived from treatment-planning CT combined with the tumour volume. It was validated with a C-Index of 0.66. Patient stratification in groups at low and high risk for loco-regional recurrence showed significant differences and calibration showed adequate results.

The prognostic performance of the final model in the present study was similar to the results of the best-performing Cox model in Leger et al. [21] that also assessed LRC in locally advanced HNSCC patients based on CT images (C-Index: 0.68). A similar performance was observed by Aerts et al. [18] (C-Index: 0.69) and Bogowicz et al. [14] (C-Index: 0.72) with their CT-derived signatures that were based on different HNSCC cohorts. While the signature from Aerts et al. [18] was shown to be highly correlated to tumour volume [40] and the signature of Bogowicz et al. [14] included a wavelet-filtered feature, we identified additional radiomics features that weakly correlated with tumour volume and did not include wavelet features, which have been found difficult to reproduce [41].

All combinations of feature selection and machine learning algorithms showed the primary tumour volume as the most-occurring clinical feature. Others, like p16 status or alcohol intake, were selected much less often. Tumour volume is a validated biomarker for overall survival in HNSCC [3–6] and was expected for radiobiological reasons [42,43], explaining its selection. In our cohorts, p16 status was not selected since it did not show a significant association with LRC, even when considering the subgroup of oropharyngeal tumours. This may in part be explained by the lower fraction of p16-positive tumours and the higher tumour volume compared to Linge et al. [3].

Two CT radiomics features were selected in the final signature. The texture feature log_ngl_hdhge may capture aspects of tumour microenvironment heterogeneity [44], associated with more recurrent tumours in HNSCC [45]. The other feature, stat_p10 represents the 10th percentile of the intensity histogram within the GTV. It was shown to weakly correlate with the proliferation index Ki67

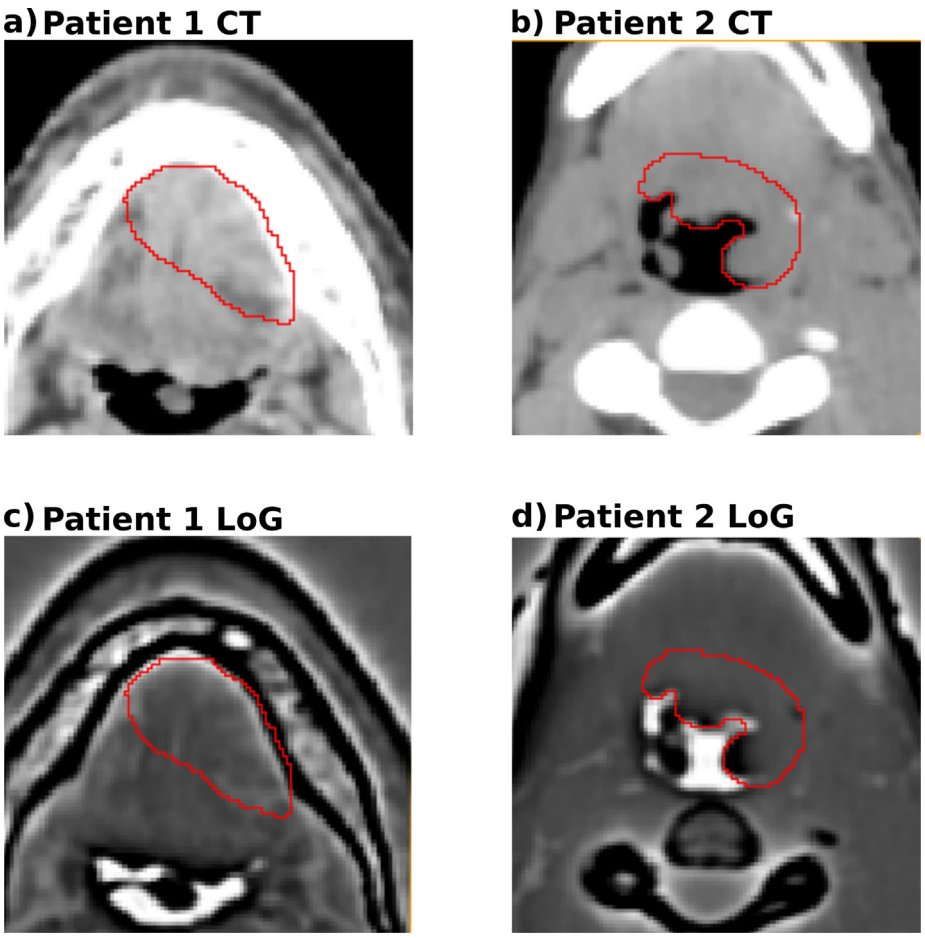

Fig. 4. Exemplary CT slices (a,b) and slices of the Logarithm of Gaussian (LoG) filtered images (c,d) of two patients in the discovery cohort (10x10 cm^2 crop). Patient 1 showed highly expressed features in the final signature (tumour volume, stat_p10 and log_ngl_hdhge) and a loco-regional recurrence developed shortly after treatment. For patient 2, features showed a lower expression and no recurrence was observed during follow-up. For patient 1, a more heterogeneous tumour can be seen, with zones of varying intensity across the slice, which is emphasised in the LoG image. For patient 2, a more uniform tumour is visible. Red contours mark the primary tumour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Table with information of features in the final clinical-radiomics model: Hazard Ratio (HR) (95% CI) along with model p-values and transformation parameters for z-transform (z-shift and z-scale) and Yeo-Johnson parameter (λ).

Feature	HR [95% CI]	p-value	z-shift	z-scale	λ
GTV (cm^3)	1.248 [0.962–1.619]	0.096	3.457	0.881	0
stat_p10	1.171 [0.983–1.396]	0.077	28.633	62.008	1.5
log_ngl_hdhge	1.215 [0.902–2.095]	0.14	10.968	0.614	0

GTV: gross tumour volume, stat_p10: 10th percentile of intensity histogram, log_ngl_hdhge: high-dependence high-emphasis of the NGL matrix in the Laplacian of Gaussian image

in HNSCC [46]. Combining clinical and radiomics features improved results compared to a clinical-only model for discrimination and stratification, as was also shown by Zhai et al. [47] for HNSCC. However, linkage of specific CT radiomics features to underlying biological mechanisms are not currently well-established and should be studied in the future.

Model performance may be increased by including additional imaging modalities: Bogowicz et al. [14] showed that a signature consisting of CT and PET features had an increased performance for prognosis of LRC (C-Index: 0.73) compared to PET alone (C-Index:0.71). Deep learning may be an interesting approach as convolutional neural networks (CNN) can learn abstract representations from images without the need of hand-crafted features. Using a previously-trained CNN (transfer learning) for LRC prognosis in HNSCC an average area under the curve (AUC) of 0.64 was achieved by Diamant et al. [48], outperforming the 0.5 reached using that same CNN by Valières et al. [49]. Haarburger et al.[50] used CNNs to classify patients with non-small cell lung cancer on a publicly available dataset. Features learned by the CNNs were then employed for hazard prediction on a Cox model (C-Index: 0.623) , showing a slightly higher performance than the signature from Aerts et al. (C-Index: 0.609) [18]. However, improved prognostic value compared to the conventional radiomics approach still must be shown.

Presently, prognostic radiomics models do not yet translate into clinical application. Firstly, feature reproducibility is an issue as there is a lack of consensual guidelines on how to extract and define radiomics features. The IBSI [25] aims to establish such a consensus and reporting guidelines for the methods employed for extraction. Secondly, there is underreporting in radiomics studies as defined by the TRIPOD statement [51], such as handling of missing data or model specifications like baseline survival, which

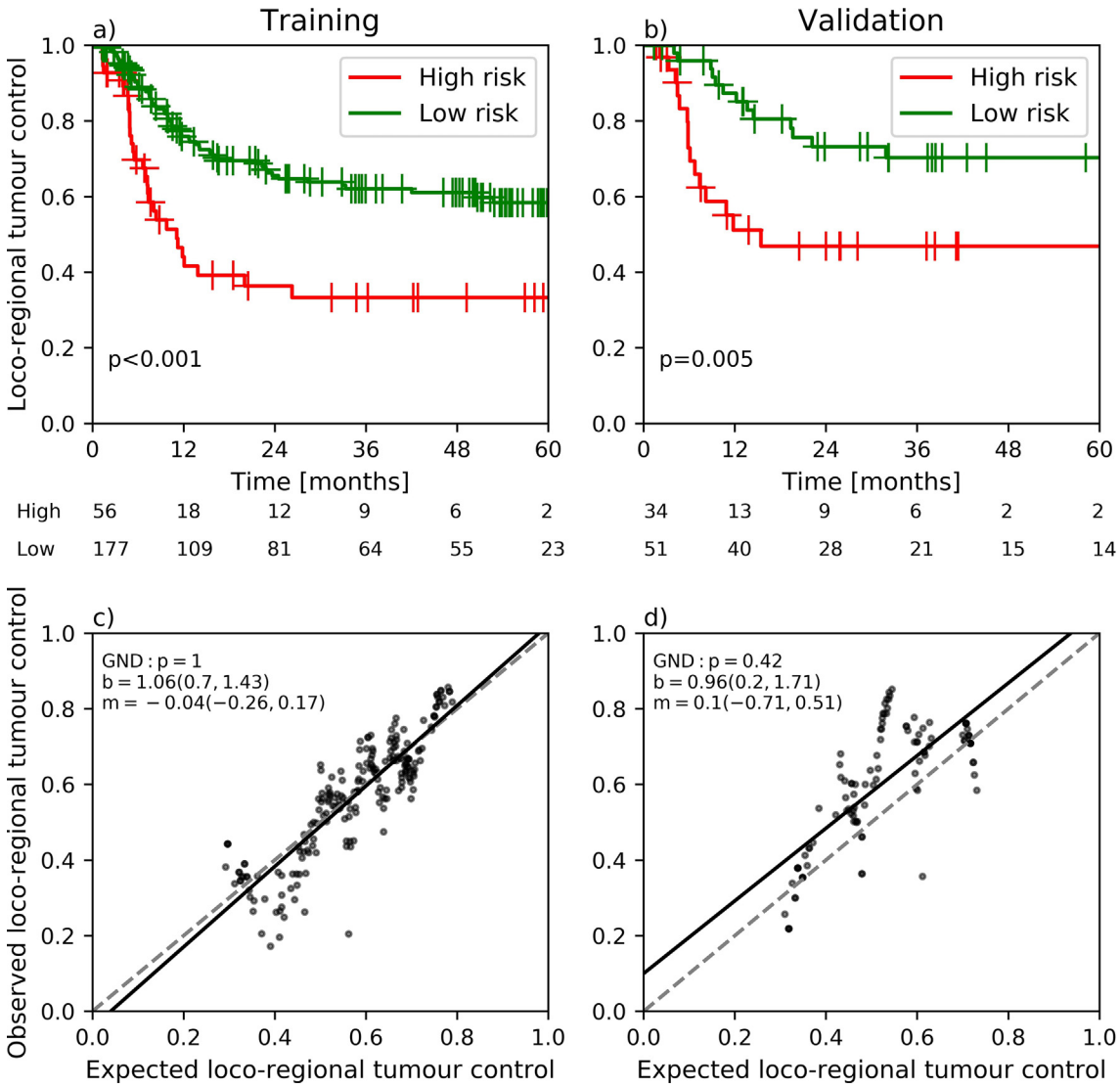

Fig. 5. Stratification and calibration for the final model in discovery and validation cohorts. Significant differences in loco-regional tumour control (LRC) were observed in (a) the discovery cohort (training) and (b) the validation cohort between low and high-risk groups as measured by the log-rank test. In calibration, expected and observed LRC were not significantly different 24 months after treatment in training (c) and validation (d) as shown by the GND test. Linear fits with slope b and intercept m are shown (with 95% confidence intervals) in comparison to the ideal diagonal dashed line.

leads to limited reproducibility of findings [52]. Furthermore, prospective studies for validation of radiomics signatures are rare, which are essential for progression towards clinical application [53]. To tackle such problems, we have established our radiomics features in accordance to the IBSI guidelines and report on the parameters and algorithms used for their extraction, transformation, stability analysis, and modelling. We use a clear end-to-end modelling strategy, optimise hyperparameters and resample data to help reduce overfitting. We also report on the results of our signature based on three clear aspects on an independent validation cohort: discrimination, stratification and calibration. Finally, we aim to apply the signature to prospective data of the HNPrädBio trial of the DKTK-ROG (www.clinicaltrials.gov, NCT02059668), which received primary RCTx.

In this study, we developed and validated a clinical-radiomics signature for assessing LRC in locally advanced HNSCC patients. This signature combined the primary tumour volume with two independent CT radiomics features. In the future, we aim to further validate the signature with data from the prospective HNPrädBio trial of the DKTK-ROG before potential application in an interventional clinical trial on dose adaptation.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2020.11.011.

References

[1] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65:87–108.
[2] Budach V, Tinhofer I. Novel prognostic clinical factors and biomarkers for outcome prediction in head and neck cancer: a systematic review. Lancet Oncol. 2019:e313–26.
[3] Linge A, Schmidt S, Lohaus F, Krenn C, Bandurska-Luque A, Platzek I, et al. Independent validation of tumour volume, cancer stem cell markers and hypoxia-associated gene expressions for HNSCC after primary radiochemotherapy. Clin Transl Radiat Oncol 2019;16:40–7.

[4] Dubben HH, Thames HD, Beck-Bornholdt HP. Tumor volume: A basic and specific response predictor in radiotherapy. Radiother Oncol. 1998;47:167–74.
[5] Soliman M, Yaromina A, Appold S, Zips D, Reiffenstuhl C, Schreiber A, et al. GTV differentially impacts locoregional control of non-small cell lung cancer (NSCLC) after different fractionation schedules: Subgroup analysis of the prospective randomized CHARTWEL trial. Radiother Oncol. 2013;106:299–304.
[6] Linge A, Lohaus F, Löck S, Nowak A, Gudziol V, Valentini C, et al. HPV status, cancer stem cell marker expression, hypoxia gene signatures and tumour volume identify good prognosis subgroups in patients with HNSCC after primary radiochemotherapy: A multicentre retrospective study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG). Radiother Oncol. 2016;121:364–73.
[7] Sivars L, Landin D, Grün N, Vlastos A, Marklund L, Nordemar S, et al. Validation of human papillomavirus as a favourable prognostic marker and analysis of CD8+ tumour-infiltrating lymphocytes and other biomarkers in cancer of unknown primary in the head and neck region. Anticancer Res. 2017;37:665–74.
[8] Krupar R, Robold K, Gaag D, Spanier G, Kreutz M, Renner K, et al. Immunologic and metabolic characteristics of HPV-negative and HPV-positive head and neck squamous cell carcinomas are strikingly different. Virchows Arch. 2014;465:299–312.
[9] Lendahl U, Lee KL, Yang H, Poellinger L. Generating specificity and diversity in the transcriptional response to hypoxia. Nat. Rev. Genet. 2009;10:821–32.
[10] Toustrup K, Sørensen BS, Nordsmark M, Busk M, Wiuf C, Alsner J, et al. Development of a hypoxia gene expression classifier with predictive impact for hypoxic modification of radiotherapy in head and neck cancer. Cancer Res. 2011;71:5923–31.
[11] Eustace A, Mani N, Span PN, Irlam JJ, Taylor J, Betts GNJ, et al. A 26-gene hypoxia signature predicts benefit from hypoxia-modifying therapy in laryngeal cancer but not bladder cancer. Clin Cancer Res. 2013;19:4879–88.
[12] Löck S, Perrin R, Seidlitz A, Bandurska-Luque A, Zschaeck S, Zöphel K, et al. Residual tumour hypoxia in head-and-neck cancer patients undergoing primary radiochemotherapy, final results of a prospective trial on repeat FMISO-PET imaging. Radiother Oncol. 2017;124:533–40.
[13] Keek SA, Leijenaar RT, Jochems A, Woodruff HC. A review on radiomics and the future of theranostics for patient selection in precision medicine. Br. J. Radiol. 2018;91:2017926.
[14] Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. Computed Tomography Radiomics Predicts HPV Status and Local Tumor Control After Definitive Radiochemotherapy in Head and Neck Squamous Cell Carcinoma. Int J Radiat Oncol Biol Phys. 2017;99:921–8.
[15] Cozzi L, Franzese C, Fogliata A, Franceschini D, Navarria P, Tomatis S, et al. Predicting survival and local control after radiochemotherapy in locally advanced head and neck cancer by means of computed tomography based radiomics. Strahlentherapie und Onkol. 2019;195:805–18.
[16] Vallieres M, Kumar A, Sultanem K, El Naqa I. FDG-PET Image-Derived Features Can Determine HPV Status in Head-and-Neck Cancer. Int J Radiat Oncol. 2013;87:S467.
[17] Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RTH, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. Oral Oncol. 2017;71:150–5.
[18] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:1–9.
[19] Leijenaar RTH, Carvalho S, Hoebers FJP, Aerts HJWL, Van Elmpt WJC, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol (Madr). 2015;54:1423–9.
[20] Leger S, Zwanenburg A, Pilz K, Zschaeck S, Zöphel K, Kotzerke J, et al. CT imaging during treatment improves radiomic models for patients with locally advanced head and neck cancer. Radiother Oncol. 2019;130:10–7.
[21] Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, et al. A comparative study of machine learning methods for time-To-event survival data for radiomics risk modelling. Sci Rep. 2017;7:13206.
[22] Zips D, Zöphel K, Abolmaali N, Perrin R, Abramyuk A, Haase R, et al. Exploratory prospective trial of hypoxia-specific PET imaging during radiochemotherapy in patients with locally advanced head-and-neck cancer. Radiother Oncol. 2012;105:21–8.
[23] Welz S, Mönnich D, Pfannenberg C, Nikolaou K, Reimold M, La Fougère C, et al. Prognostic value of dynamic hypoxia PET in head and neck cancer: Results from a planned interim analysis of a randomized phase II hypoxia-image guided dose escalation trial. Radiother Oncol. 2017;124:526–32.
[24] Depeursinge A, Andrearczyk V, Whybra P, van Griethuysen J, Müller H, Schaer R, et al. Standardised convolutional filtering for radiomics. 2020 Available from: http://arxiv.org/abs/2006.05470.
[25] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295:328–38.
[26] Zwanenburg A, Leger S, Starke S. GitHub - oncoray/mirp: Medical Image Radiomics Processor. Available from: https://github.com/oncoray/mirp.
[27] Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019;9:614.
[28] Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15:155–63.
[29] Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal. 2009;53:3735–45.
[30] Yeo INK, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. Oxford University Press 2000;87:954–9.
[31] Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226–38.
[32] Tibshirani R. The lasso method for variable selection in the cox model. Stat Med. 1997;16:385–95.
[33] Bühlmann P, Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. Stat Sci. 2007;22:477–505.
[34] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann. Appl Stat. 2008;2:841–60.
[35] Hutter F, Hoos HH, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2011.
[36] Meinshausen N, Bühlmann P. Stability Selection. J R Stat Soc Ser B Stat Methodol. 2010;72:417–73.
[37] DiCiccio TJ, Efron B. Bootstrap confidence intervals. Stat Sci. 1996;11:189–212.
[38] Hothorn T, Lausen B. On the exact distribution of maximally selected rank statistics. Comput Stat Data Anal. 2003;43:121–37.
[39] Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. Stat Med. 2015;34:1659–80.
[40] Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: The need for safeguards. Radiother Oncol. 2019;130:2–9.
[41] Bogowicz M, Leijenaar RTH, Tanadini-Lang S, Riesterer O, Pruschy M, Studer G, et al. Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models. Radiother Oncol. 2017;125:385–91.
[42] Baumann M, DuBois W, Suit HD. Response of human squamous cell carcinoma xenografts of different sizes to irradiation: Relationship of clonogenic cells, cellular radiation sensitivity in vivo, and tumor rescuing units. Radiat Res. 1990;123:325–30.
[43] Baumann M, Krause M, Hill R. Exploring the role of cancer stem cells in radioresistance. Nat. Rev. Cancer. 2008;7:545–54.
[44] Katsoulakis E, Yu Y, Apte AP, Leeman JE, Katabi N, Morris L, et al. Radiomic analysis identifies tumor subtypes associated with distinct molecular and microenvironmental factors in head and neck squamous cell carcinoma. Oral Oncol. 2020;110:104877.
[45] Choi JW, Lee D, Hyun SH, Han M, Kim JH, Lee SJ. Intratumoural heterogeneity measured using FDG PET and MRI is associated with tumour-stroma ratio and clinical outcome in head and neck squamous cell carcinoma. Clin Radiol. W.B. 2017;72:482–9.
[46] Meyer HJ, Hamerla G, Höhn AK, Surov A. CT texture analysis-correlations with histopathology parameters in head and neck squamous cell carcinomas. Front Oncol. 2019;9:444.
[47] Zhai TT, Langendijk JA, van Dijk LV, Halmos GB, Witjes MJH, Oosting SF, et al. The prognostic value of CT-based image-biomarkers for head and neck cancer patients treated with definitive (chemo-)radiation. Oral Oncol. 2019;95:178–86.
[48] Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Sci Rep. 2019;9:2764.
[49] Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJWL, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7:10117.
[50] Haarburger C, Weitz P, Rippel O, Merhof D. Image-based survival prediction for lung cancer patients using CNNS. Proc - Int Symp Biomed. Imaging. 2019:1197–201.
[51] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ. 2015;350.
[52] Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020;30:523–36.
[53] Guha A, Connor S, Anjari M, Naik H, Siddiqui M, Cook G, et al. Radiomic analysis for response assessment in advanced head and neck cancers, a distant dream or an inevitable reality? A systematic review of the current level of evidence. Br. J. Radiol. 2020;93:20190496.