Radiomics prognostic analysis of PET/CT images in a multicenter head and neck cancer cohort: investigating ComBat strategies, sub-volume characterization, and automatic segmentation

Hui Xu, Nassib Abdallah, Jean-Marie Marion, Pierre Chauvet, Clovis Tauber, Thomas Carlier, Lijun Lu, Mathieu Hatt

European Journal of Nuclear Medicine and Molecular Imaging (2023) 50:1720–1734
https://doi.org/10.1007/s00259-023-06118-2

ORIGINAL ARTICLE

Radiomics prognostic analysis of PET/CT images in a multicenter head and neck cancer cohort: investigating ComBat strategies, sub-volume characterization, and automatic segmentation

Hui Xu[1,2] · Nassib Abdallah[2] · Jean-Marie Marion[3] · Pierre Chauvet[3] · Clovis Tauber[4] · Thomas Carlier[5] · Lijun Lu[1,6] · Mathieu Hatt[2]

Received: 20 October 2022 / Accepted: 16 January 2023 / Published online: 24 January 2023

Abstract

Purpose This study aimed to investigate the impact of several ComBat harmonization strategies, intra-tumoral sub-volume characterization, and automatic segmentations for progression-free survival (PFS) prediction through radiomics modeling for patients with head and neck cancer (HNC) in PET/CT images.

Methods The HECKTOR MICCAI 2021 challenge set containing PET/CT images and clinical data of 325 oropharynx HNC patients was exploited. A total of 346 IBSI-compliant radiomic features were extracted for each patient's primary tumor volume defined by the reference manual contours. Modeling relied on least absolute shrinkage Cox regression (Lasso-Cox) for feature selection (FS) and Cox proportional-hazards (CoxPH) models were built to predict PFS. Within this methodological framework, 8 different strategies for ComBat harmonization were compared, including before or after FS, in feature groups separately or all features directly, and with center or clustering-determined labels. Features extracted from tumor sub-volume clustering were also investigated for their prognostic additional value. Finally, 3 automatic segmentations (2 threshold-based and a 3D U-Net) were also compared. All results were evaluated with the concordance index (*C*-index).

Results Radiomics features without harmonization, combined with clinical factors, led to models with *C*-index values of 0.69 in the testing set. The best version of ComBat harmonization, i.e., after FS, for feature groups separately and relying on clustering-determined labels, achieved a *C*-index of 0.71. The use of features extracted from tumor sub-volumes further improved the *C*-index to 0.72. Models that relied on the automatic segmentations yielded close but slightly lower prognostic performance (0.67–0.70) compared to reference contours.

Conclusion A standard radiomics pipeline allowed for prediction of PFS in a multicenter HNC cohort. Applying a specific strategy of ComBat harmonization improved the performance. The extraction of intra-tumoral sub-volume features and automatic segmentation could contribute to the improvement and automation of prognosis modeling, respectively.

Keywords PET/CT · Radiomics · ComBat · Sub-volume · Automatic segmentation

This article is part of the Topical Collection on Oncology—Head and Neck.

✉ Lijun Lu
ljlubme@gmail.com

[1] School of Biomedical Engineering and Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, 1023 Shatai Road, Guangzhou 510515, Guangdong, China

[2] LaTIM, INSERM, UMR 1101, University Brest, Brest, France

[3] LARIS, Univ Angers, EA7315 Angers, France

[4] INSERM U930, Université François Rabelais de Tours, Tours, France

[5] Nuclear Medicine Department, CHU and CRCINA, INSERM, CNRS, Univ Angers, Univ Nantes, Nantes, France

[6] Pazhou Lab, Guangzhou 510330, China

Introduction

Head and neck cancer (HNC) is the seventh most common cancer worldwide [1]. Early prognosis/outcome prediction would be beneficial for tailoring individualized treatment strategies. [^{18}F]FDG PET/CT is a powerful tool in managing HNC patients, thanks to the complementary nature of both anatomical and functional information [2]. Quantitative analysis by extracting high-dimensional features from medical images, known as radiomics, has shown some potential in improving clinical decision-making [3]. However, it still suffers from several limitations and is yet to translate to clinical routine. These limitations include (i) insufficient automation, (ii) issues with multicentric data harmonization, (iii) modeling unresolved issues, and (iv) the "black box effect," i.e., the lack of interpretability of the resulting multiparametric models [4–6]. Although the field of radiomics is currently quickly evolving toward the use of deep learning methods, the standard radiomics workflow relying on the extraction of handcrafted features from a delineated tumor volume still has potential value, especially for the easier explainability of the resulting models and the capability to learn on limited size datasets [7].

Multicenter modeling is crucial in facilitating the clinical translation of radiomics, because it can allow producing high level of clinical proof regarding their added value with respect to the usual clinical factors relied upon for clinical decision. However, it is a challenging task because the image properties (e.g., intensity, spatial resolution, textures) are affected by the variations of imaging devices, acquisition protocols, and reconstruction algorithms [8]. First, standardizing the procedures of image acquisition and quantitative analysis, such as following the EANM EARL initiative for PET/CT imaging [9], is recommended for multicentric clinical trials; however, it can only be exploited in prospective data collection and has been shown to be insufficient [10]. Moreover, some post-processing techniques (i.e., intensity mapping [11], generative adversarial networks [12]) have been used to harmonize images. Instead of processing images, statistical harmonization of features (e.g., ComBat [13]) is widely used for multicenter dataset, with the advantage of not requiring feature re-extraction and being applicable to retrospectively collected data [14, 15]. Still, 18% of studies using ComBat reported no benefit after harmonization [16]. Thus, more efficient versions of ComBat for radiomics feature harmonization need to be further explored [17].

Standard engineered radiomic features have been used extensively in previous studies. They are typically calculated across the entire tumor 3D volume or a selected 2D slice. This approach may not be sufficient to characterize intra-tumoral regional variations and capture the underlying biological process [18]. A finer quantification of intra-tumoral spatial heterogeneity in sub-volume level was reported with potential to better predict outcome [19, 20]. Moreover, a vast majority of current radiomics studies are based on semi-automatic or even fully manual contours of tumor volume, which is time-consuming and error-prone and hinders the processing of very large datasets [5, 21]. It is thus useful to investigate the application of faster and more reproducible automatic segmentation toward building fully automated radiomics pipelines.

Overall, this study had 3 main objectives within the context of prognosis prediction for HNC patients through a radiomics pipeline, namely to investigate (i) various strategies for ComBat harmonization of radiomic features, (ii) complementary value of intra-tumoral sub-volume characterization beyond classical radiomics features, and (iii) impact of relying on automatic segmentation. We decided to carry out these analyses in a publicly available dataset containing a large, well-curated, and annotated multicentric HNC cohort. This dataset was made available in the HEad and neCK TumOR segmentation and outcome prediction challenge (HECKTOR) organized in 2021, hosted by the International Conference on Medical Image Computing and Computer Assisted Intervention[1] (MICCAI) [22]. This ensures a higher reproducibility of our work, and it allows a comparison of our performance with those obtained by challengers in the 2021 edition.

Materials and methods

The overall study design is shown in Fig. 1.

Dataset

This study used the entire dataset from the HECKTOR 2021 challenge, which contains 325 patients (224 training and 101 testing) from six centers (Table 1) with pre-treatment [^{18}F]FDG PET/CT images (imaging protocols in Supplementary Material E1), manually (by clinical experts) delineated gross tumor volume (GTV), bounding box coordinates locating the oropharynx region containing the tumor, and clinical information. The data from one center, CHUP (Poitiers, France), were split into the training and testing sets. The predicted endpoint is progression-free survival (PFS) and the evaluation metric is the concordance index (*C*-index). Clinical data without missing values is provided in Table 2. The dataset is available upon request through the challenge website.[2]

[1] www.miccai2021.org.

[2] https://www.aicrowd.com/challenges/miccai-2021-hecktor.

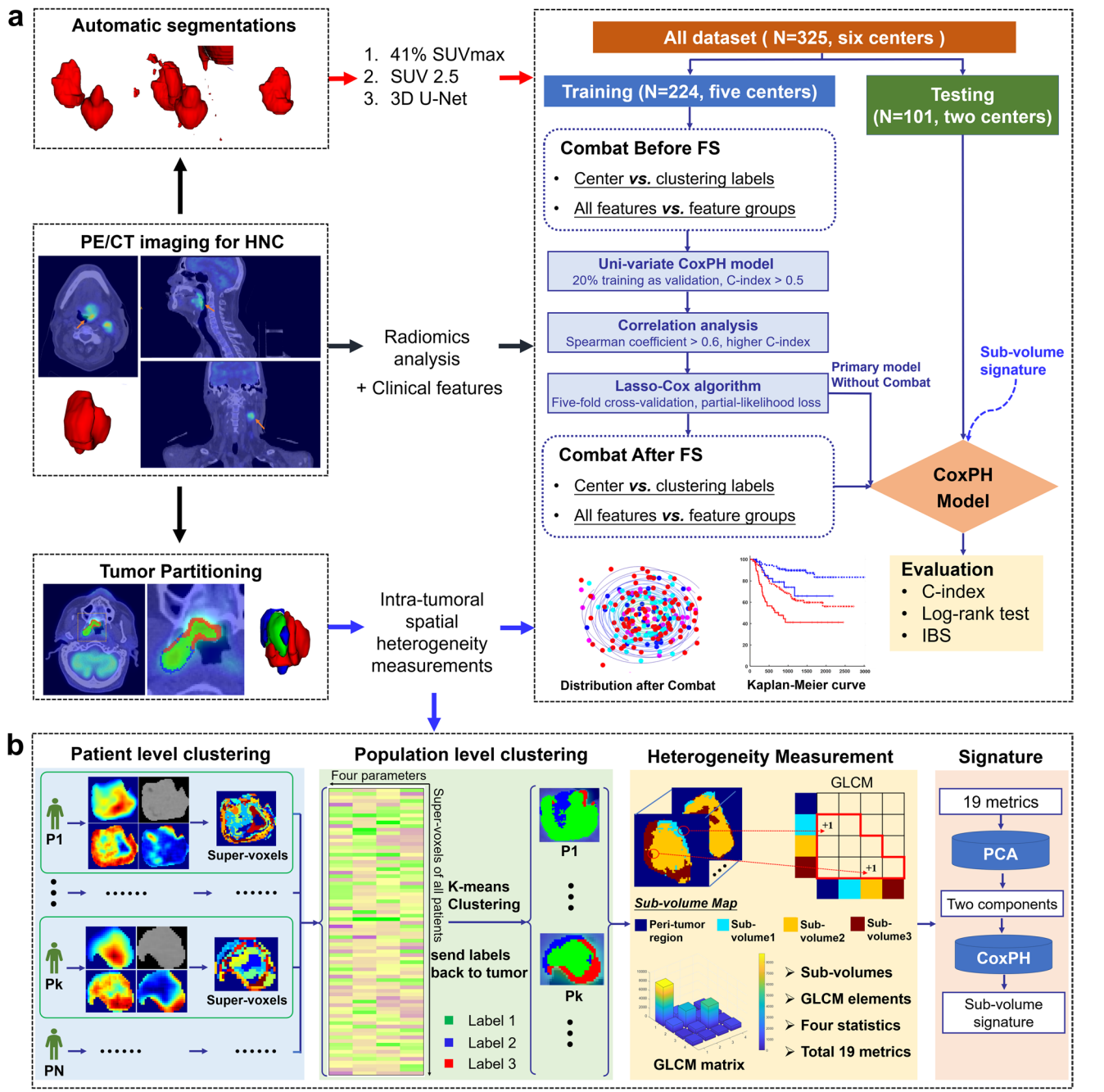

Fig. 1 Overview of the study design and the radiomics pipeline (**a**). Detailed illustration for the characterization of tumor sub-volumes (**b**)

Table 1 The statistics of multicenter datasets

Center	Device	No. of patients	Set
HGJ	Discovery ST, GE Healthcare	55	Training
CHUS	Gemini GXL 16, Philips	72	Training
HMR	Discovery STE, GE Healthcare	18	Training
CHUM	Discovery STE, GE Healthcare	56	Training
CHUP	Biography mCT 40 ToF, Siemens	23/48	Training/testing
CHUV	Discovery D690 ToF, GE Healthcare	53	Testing

Table 2 Clinical information of HNC cohorts

Characteristic	All ($N=325$)	Training (N = 224)	Testing (N = 101)	p
Age, year				
Median (range)	61 (34–90)	63 (34–90)	62 (40–84)	0.176
Mean ± SD	62.3 ± 9.4	62.8 ± 9.5	61.2 ± 9.1	
Gender, no. (%)				
Male	249 (76.6%)	167 (74.6%)	82 (81.2%)	0.191
Female	76 (23.4%)	57 (25.5%)	19 (18.8%)	
T stage, no. (%)				
T1	34 (10.5%)	26 (11.6%)	8 (7.9%)	0.063
T2	121 (37.2%)	94 (42.0%)	27 (26.7%)	
T3	102 (31.4%)	58 (25.9%)	44 (43.6%)	
T4	68 (20.9%)	46 (20.5%)	22 (21.8%)	
N stage, no. (%)				
N0	45 (13.8%)	33 (14.7%)	12 (11.9%)	0.101
N1	46 (14.2%)	26 (11.6%)	20 (19.8%)	
N2	208 (64.0%)	150 (67.0%)	58 (57.4%)	
N3	26 (8.0%)	15 (6.7%)	11 (10.9%)	
M stage, no. (%)				
M0	316 (97.2%)	220 (98.2%)	96 (95.1%)	0.108
M1	9 (2.8%)	4 (1.8%)	5 (4.9%)	
TNM stage				
I	8 (2.4%)	4 (1.8%)	4 (4.0%)	0.306
II	26 (8.0%)	19 (8.5%)	7 (6.9%)	
III	48 (14.8%)	29 (13.0%)	19 (18.8%)	
IV	243 (74.8%)	172 (76.8%)	71 (70.3%)	
Treatment				
Radiotherapy	37 (11.4%)	27 (12.1%)	10 (9.9%)	0.572
Chemo-radiotherapy	288 (88.6%)	197 (87.9%)	91 (90.1%)	
PFS, days				
Median (range)	999 (37–3067)	1170 (106–3067)	596 (37–2451)	< 0.05
Mean ± SD	1067 ± 626	1224 ± 602	721 ± 536	
Progression	96 (29.5%)	56 (25.0%)	40 (39.6%)	0.072
Censoring	229 (70.5%)	168 (75.0%)	61 (60.4%)	

Pre-processing and feature extraction

PET/CT images were acquired by the various scanners used in each center (Table 1). Details of imaging protocol and GTV delineation are provided in [22]. All CT scans were acquired without contrast agent. PET images were converted to standardized uptake values (SUV). PET images were resampled to the same resolution as CT images by linear interpolation for utilizing the provided GTV contours. Radiomics features were extracted from both PET and CT images using the open-source Imaging Biomarker Standardization Initiative (IBSI) compliant package of Standardized Environment for Radiomics Analysis (SERA) [23]. In total, 346 features were calculated from the provided GTV contours (or the mask of automatic segmentations, see "Impact of automatic segmentation") for each patient after resampling images to isotropic voxel of $2 \times 2 \times 2$ mm^3 through linear interpolation and a fixed bin number (FBN) discretization with 64 bins. Radiomics features are summarized in Supplementary Table S1.

Prognosis analysis

Radiomics model construction

Feature selection (FS) and model building (Fig. 1a) exploited the training set only to maintain the independence of external testing. Twenty percent of the training data was randomly selected as an internal validation set for FS. First, the non-meaningful features with variance of zero were removed, and Z-score normalization was performed for each feature. Second, features were ranked according to the C-index in the univariate Cox proportional-hazards (CoxPH) model, and those with a C-index equal to 0.50 or lower were removed.

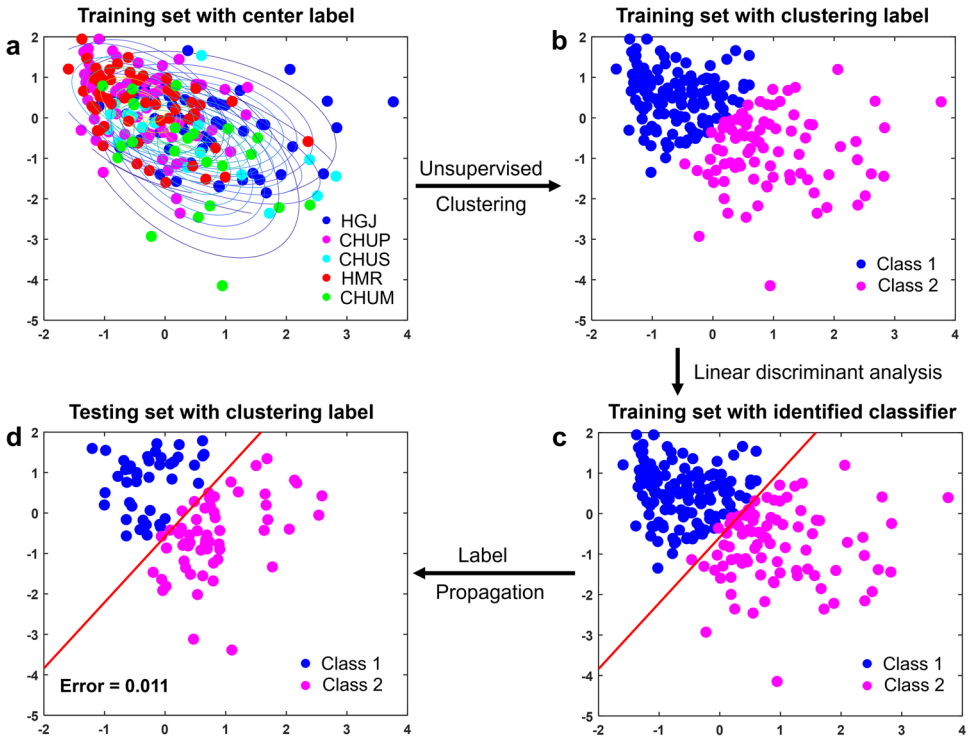

Fig. 2 The process of clustering label identification. K-means clustering algorithm identifies the clustering labels of the training set (**a**, **b**), while label propagation is used to obtain the labels of testing set (**c**, **d**)

Spearman rank correlation analysis was then relied upon to identify redundant features. For features with a correlation of 0.60 or above, the feature with the lower *C*-index was removed. The optimal subset was ultimately determined by the least absolute shrinkage and selection operator (LASSO) Cox regression algorithm, which was adopted to construct multivariate CoxPH models for PFS prediction. Associations between features and PFS were reported as hazard ratio (HR) and 95% confidence intervals (CIs).

ComBat harmonization strategies

ComBat is a data-driven method initially developed to correct for the batch-effect in genomics [13], and then used in radiomics studies to harmonize multicentric data [14, 15]. We explored several strategies by applying ComBat before ("B") or after ("A") FS, to all features ("A") or to different feature groups separately ("G"), and with center ("C") or auto-derived labels through unsupervised K-means clustering ("K"). This led to eight models denoted BAC, BAK, BGC, BGK, AAC, AAK, AGC, and AGK. For models AGC and AGK, one and two features respectively were eliminated due to their correlation of more than 0.6 (criterion of FS); then, the features were replaced by the corresponding primary features without harmonization to generate two derived models, namely AGC* and AGK*. Thus, 10 ComBat models were investigated in this study. Clustering labels of the training set were obtained by a K-means clustering algorithm. The number of clusters was determined in the range [2, 10] by the optimal Silhouette score. Clustering labels of the testing set were generated through label propagation by linear discriminant analysis (Fig. 2), which maintained the independence of testing. The one-way ANOVA test with a significant level set as $p < 0.001$ was adopted to split features into different groups which were affected by center-effect to different degrees (Fig. 3). Non-parametric estimation was used for ComBat harmonization, and no covariate was introduced.

Hybrid clinical-radiomics model construction

Feature selection for clinical features was performed by univariate CoxPH models together with correlation analysis with the same threshold parameters described above. Hybrid clinical-radiomics models were constructed by combining the selected clinical and radiomics features via multivariate CoxPH models.

Tumor sub-volume characterization

The characterization of intra-tumoral spatial heterogeneity in sub-volumes of the tumor consisted of three steps (Fig. 1b). First, a two-stage voxel clustering performed at both individual and population levels was used to split each tumor into 3 non-overlapping sub-volumes, taking PET, CT, and their corresponding local entropy-filtered images as inputs (details in Supplementary Material E2). Second, based on the generated sub-volume maps, 19 quantitative metrics

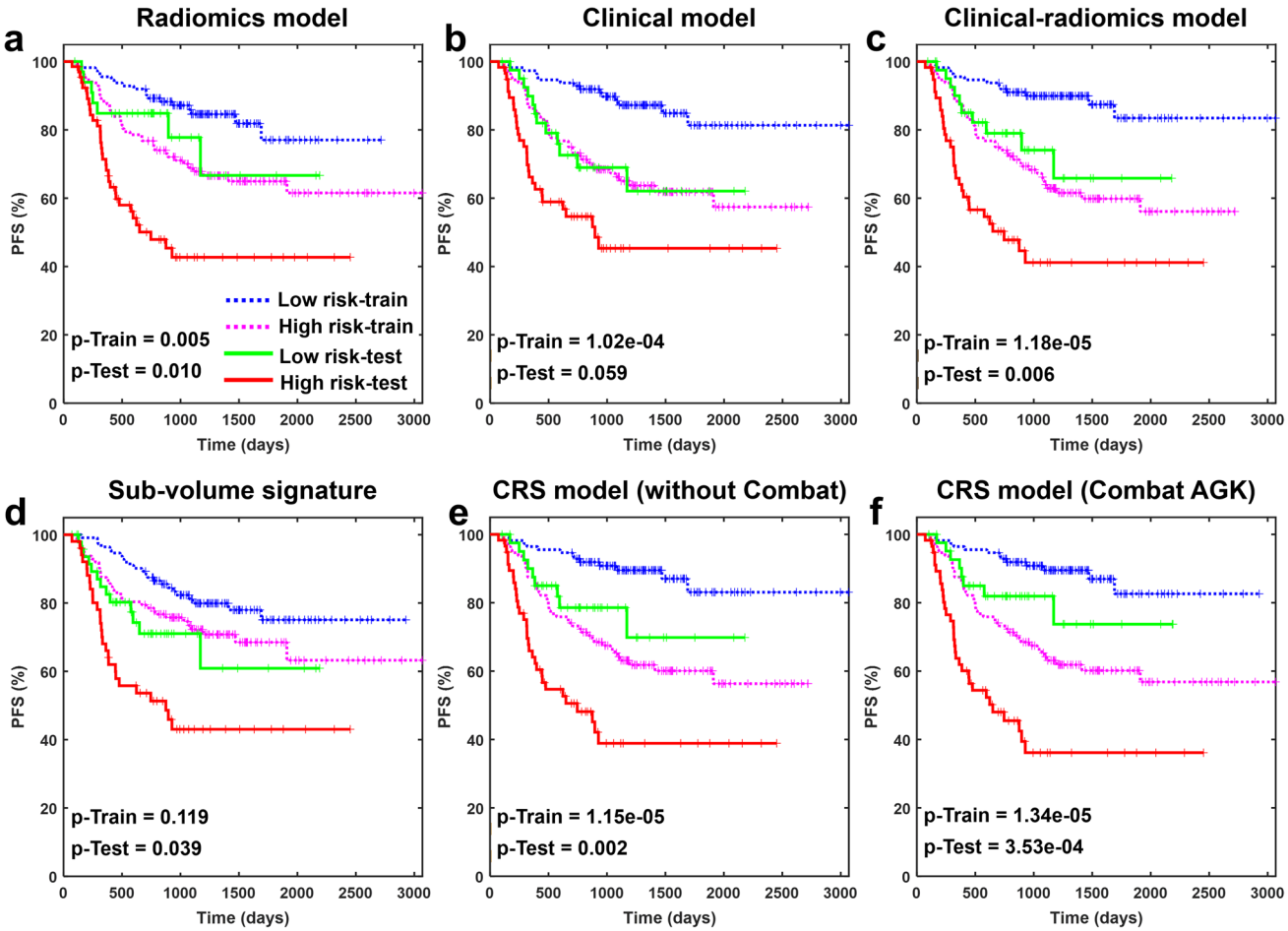

Fig. 3 Kaplan–Meier curves for radiomics (**a**), clinical (**b**), and clinical-radiomics (**c**) model in the training and testing set, respectively. Kaplan–Meier curves for the sub-volume signature (**d**), clinical-radiomics model combined with the sub-volume signature (CRS model) without ComBat harmonization (**e**), and CRS model with ComBat strategy of AGK (**f**)

were calculated for each patient to depict their spatial relationships (see Supplementary Table S2). Absolute counts and the physical volume of each sub-volume were firstly extracted (6 metrics). Then, gray-level co-occurrence matrix (GLCM) was constructed to depict the spatial interactions of different sub-volumes and peri-tumoral region. Here, three diagonal elements represented the connected size of each sub-volume; six off-diagonal elements represented the size of border where different sub-volumes intersect (9 metrics). Moreover, 4 statistical metrics of GLCM, namely contrast, correlation, energy, and homogeneity, were extracted. Third, principal component analysis (PCA) was applied to reduce the dimensionality of these metrics due to their high correlations. The mean and principal component coefficients of the training set were used to transform the testing set into the same domain. Two uncorrelated principal components were extracted and concatenated into one signature by the CoxPH model to characterize intra-tumoral spatial heterogeneity. The complementary value in predicting PFS of the sub-volume signature beyond the classical clinical-radiomics model was evaluated.

Impact of automatic segmentation

The entire workflow described above was first applied by relying on the provided GTVs, which were manually contoured by experts. To investigate whether the manual delineation could be replaced by a fully automated pipeline, we implemented three fully automatic segmentation methods and then repeated the entire pipeline described above, using the different segmented volumes as inputs. A relative threshold of 41% of SUV_{max} and a fixed threshold of $SUV_{above2.5}$ were separately applied to the provided bounding box in the PET image to extract the tumor volume (the corresponding volume was replicated on the CT image). Additionally, a deep convolutional neural network based on 3D U-Net architecture [24] was trained for fully automatic tumor segmentation with both PET and CT modalities as inputs (see

Table 3 Univariate and multivariate analysis of clinical, radiomics, and their combined model, respectively

Characteristics	Univariate analysis		Multivariate analysis			
	Univariate model		Clinical/radiomics model		Clinical-radiomics model	
	HR (95%CI)	*C*-index	HR (95%CI)	*C*-index	HR (95%CI)	*C*-index
Age	1.01 (0.97–1.04)	0.524	0.97 (0.51–1.86)	Training: 0.671 (0.596–0.715)	1.13 (0.85–1.49)	
Gender	1.21 (0.62–2.40)	0.527	1.02 (0.99–1.05)	Testing: 0.654 (0.599–0.701)	0.82 (0.62–1.08)	
T stage	1.66 (1.22–2.29)	0.609	1.57 (1.15–2.13)		1.11 (0.82–1.50)	
N stage	1.07 (0.72–1.59)	0.561	1.08 (0.76–1.54)		1.02 (0.74–1.42)	
M stage	41.6 (8.27–209)	0.592	29.5 (8.13–107)		1.41 (1.01–1.98)	Training: 0.694 (0.628–0.738)
TNM stage	1.11(0.72–1.71)	0.545	NA			
Treatment	1.24 (0.48–3.21)	0.522	NA			Testing: 0.690 (0.623–0.729)
PET_IS_range	1.37 (1.05–1.78)	0.649	1.28 (0.99–1.66)	Training: 0.630 (0.573–0.686)	34.3 (8.67–136)	
CT_Morph_admvee	0.87 (0.65–1.15)	0.630	0.81 (0.61–1.06)	Testing: 0.649 (0.590–0.670)	1.11 (0.77–1.60)	
CT_LI_local_peak	1.30 (0.99–1.72)	0.544	1.06 (0.82–1.39)		1.02 (0.99–1.05)	
CT_DZM_sde	0.86 (0.64–1.16)	0.515	0.89 (0.65–1.20)		0.99 (0.51–1.90)	

Abbreviation: Morph, morphology; *IS*, intensity-based statistics; *LI*, local intensity; *DZM*, gray-level distance zone matrix; *admvee*, area density (minimum volume enclosing ellipsoid); *sde*, small distance emphasis; *HR*, hazard ratio; *CI*, confidence interval; *NA*, not applicable. The 95% CIs of HR and *C*-index are shown in parentheses

Supplementary Material E3). The Dice coefficient was used to assess the overlap between the provided GTV and the resulting automated contours.

Statistics and model evaluation

Differences in clinical characteristics between the training and testing sets were assessed by the unpaired *t*-test or chi-square test (Table 2). In-Group Proportion (IGP) statistic [25] was used to measure the reproducibility of tumor sub-volume partitioning. The differences of imaging characteristics among three sub-volumes were measured by the ANOVA test. All models were evaluated on the external testing set using the *C*-index, integrated Brier score (IBS). Further evaluation consisted in selecting the median value of the prognostic signature in the training set and using it to stratify patients of the testing set into high- and low-risk groups, evaluating the difference by the log-rank test and Kaplan–Meier curves. For each model with or without harmonization, all training data were bootstrapped with 1000 repetitions to obtain the confidence intervals of each assessment metric [26], and tested on one independent testing set. *C*-index value differences between manual and automatic segmentation were evaluated by the paired non-parametric *t*-test. We also compared our results with the best ranked challengers in MICCAI HECKTOR 2021 challenge Task 2 (i.e., the reference contours not available) and Task 3 (i.e., with the reference contours available) [22]. The corrected *p*-value by Bonferroni correction below 0.05 was considered statistically significant. This study followed the guideline of Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) [27], and was evaluated using the Quality Radiomics Score.[3] A completed checklist and score table are provided in Supplementary Tables S3 and S4.

Results

Patients' characteristics

The clinical characteristics of the 325 HNC patients are summarized in Table 2. There were no significant differences between the training and testing sets for all characteristics (corrected $p = 0.063$–0.572) except for the PFS ($p < 0.01$).

Performance of prognostic models

With no harmonization, four radiomic features (PET_IS_range, CT_Morph_admvee, CT_LI_local_peak, CT_DZM_sde) were ultimately selected to construct a radiomics model with a *C*-index of 0.649 in testing (Table 3). Five

[3] https://www.radiomics.world/rqs.

Table 4 Prognostic performance of the models with different harmonization strategies

ComBat strategy	Before FS						After FS				
	All features		Feature groups		All features		Feature groups				
		Center	K-cluster	Center	K-cluster	Center	K-cluster	Center	Center	K-cluster	K-cluster
C-index	W/O	BAC	BAK	BGC	BGK	AAC	AAK	AGC	AGC*	AGK	AGK*
FS (/346)	4	4	3	1	3	4	4	3	4	2	3
R-Training	0.630	0.570	0.618	0.581	0.612	0.612	0.635	0.665	0.662	0.648	0.623
R-Testing	0.649	0.544	0.548	0.602	0.632	0.567	0.646	0.664	0.663	0.656	0.664
CR-Training	0.694	0.680	0.687	0.674	0.691	0.686	0.695	0.708	0.710	0.690	0.692
CR-Testing	0.690	0.669	0.688	0.664	0.680	0.663	0.692	0.697	0.697	0.713	0.711
CRS-Training	0.700	0.686	0.693	0.679	0.701	0.699	0.700	0.709	0.715	0.696	0.698
CRS-Testing	0.699	0.679	0.699	0.668	0.691	0.676	0.702	0.700	0.707	0.718	0.719
p-value of log-rank test											
R-Training	0.005	0.960	0.041	0.076	0.201	0.332	0.032	0.002	0.002	0.004	0.013
R-Testing	0.010	0.044	0.799	0.508	0.192	0.061	0.021	0.070	0.096	0.074	0.009
CR-Training	1.18e − 05	4.87e − 05	1.13e − 05	4.67e − 04	3.74e − 05	4.13e − 04	1.34e − 05	7.40e − 07	1.53e − 05	1.62e − 05	3.20e − 06
CR-Testing	0.006	0.126	0.002	0.112	0.030	0.127	0.006	0.021	0.005	0.001	0.003
CRS-Training	1.15e − 05	3.17e − 05	1.27e − 04	1.26e − 04	1.08e − 05	1.06e − 04	4.68e − 05	5.30e − 07	3.30e − 06	1.34e − 05	1.11e − 05
CRS-Testing	0.002	0.062	5.13e-04	0.034	0.024	0.052	0.001	0.007	0.004	3.53e-04	7.20e-04
IBS											
R-Training	0.164	0.163	0.164	0.164	0.164	0.164	0.163	0.164	0.163	0.164	0.164
R-Testing	0.227	0.248	0.237	0.240	0.233	0.242	0.226	0.219	0.217	0.222	0.222
CR-Training	0.150	0.149	0.149	0.148	0.149	0.149	0.149	0.150	0.150	0.150	0.149
CR-Testing	0.198	0.205	0.198	0.206	0.202	0.207	0.198	0.194	0.193	0.194	0.194
CRS-Training	0.146	0.145	0.146	0.145	0.145	0.145	0.146	0.147	0.146	0.146	0.146
CRS-Testing	0.196	0.203	0.197	0.204	0.200	0.204	0.196	0.193	0.191	0.193	0.193

Abbreviation: FS (/346), the number of features selected and used in the models; *W/O*, without ComBat harmonization; *R*, radiomics model; *CR*, clinical-radiomics model; *CRS*, clinical-radiomics model combined with the sub-volume signature; *K-cluster*, the labels for ComBat generated from K-means clustering algorithm. The corresponding confidence intervals are supplemented in Table S5

clinical features (age, gender, and T, N, and M stages) were selected in a clinical model achieving a *C*-index of 0.654 in testing. The hybrid clinical-radiomics model yielded higher performance with a *C*-index of 0.690. Details of model building and correlation analysis of features are provided in Supplementary Material E4. These models could significantly distinguish patients between high and low survival risk (log-rank test, $p < 0.05$), except for the clinical model (log-rank test, $p = 0.057$). Kaplan–Meier curves of radiomics, clinical, and the hybrid model are displayed in Fig. 3a to c.

Effect of ComBat harmonization

Table 4 provides the overall results of 11 models including one model without harmonization and ten models with different ComBat strategies using three evaluated metrics (*C*-index, log-rank test, and IBS) in both training and testing sets, and their confident intervals are provided in Table S5. Each strategy under comparison is provided for radiomics (R), clinical-radiomics (CR), and clinical-radiomics with sub-volume signature (CRS) models, respectively. The number of included features after FS is listed in Table 4. Based on K-means clustering, two clusters were identified as optimum to take as batches for three models (BAK, BGK, and AAK), while seven and two clusters were identified in model AGK for feature group with $p \geq 0.001$ and $p < 0.001$, respectively (Supplementary Fig. S1). An example of feature distribution before and after ComBat in model AGK was plotted in Fig. 4 and Supplementary Fig. S2.

Overall, the variability of *C*-index values on the testing set across all strategies and models under comparison was relatively large, with values ranging between 0.544 and 0.719, a median of 0.679, and a mean of 0.668 ± 0.045 A similar observation can be made on the IBS from 0.191 to 0.248 with a median of 0.203 and a mean of 0.209 ± 0.017. Harmonization led to unchanged or slightly decreased prognostic performance of models, except for a few strategies, namely after FS (rather than before), harmonizing by feature groups (instead of harmonizing all types of features together), and using labels automatically obtained through clustering rather than the center labels. The best performance of the clinical-radiomics model was achieved by model AGK with the *C*-index of 0.713 in testing compared to 0.690 without harmonization (Table 4). A similar improvement was observed with the model further combining sub-volume signature (model AGK* with *C*-index 0.719 vs. 0.699 without harmonization). Note that although these improvements can be considered small in absolute terms, they could mean moving up or down several ranks in the HECKTOR challenge ranking (see next sections for comparison).

Sub-volume characterization

Three intra-tumoral sub-volumes were consistently identified in both training and testing sets, with IGP values of 0.89, 1.00, and 0.97. Four imaging inputs in each sub-volume showed consistent distribution between the training (Fig. 5a) and testing sets (Supplementary Fig. S3). As displayed by Fig. 5b, we observed that sub-volume 1 (green label) was associated with the highest PET SUV and local entropy values, and is usually located in the core of the tumor. Sub-volume 3 (red label) had moderate PET SUV and high PET and CT entropy values and is usually located at the edges of the tumor, whereas sub-volume 2 (blue label) is usually located at the contralateral border of the tumor with the lowest PET SUV and local entropy.

Individual performances of 19 metrics are provided in Supplementary Table S6. The first two principal components that depicted a total variance of 83.3% and 14.6% were extracted to build the sub-volume signature. This signature showed low correlations with clinical features (Spearman rank correlation < 0.5), and potential prognostic power (*C*-index, 0.602 in testing set). The median value of the sub-volume signature determined in the training set led to significant risk stratification for testing patients (HR = 2.02, 95% CI: 1.09–3.77, log-rank $p = 0.039$; Fig. 3d). We observed that the predictive power of the established clinical-radiomics model was consistently improved after combining sub-volume signature (Fig. 3e, f), whether it used ComBat harmonization or not (*C*-index 0.663–0.713 vs. 0.676–0.719 in testing; Table 4).

Impact of alternative contours through automatic segmentation

The U-Net had a good performance for tumor segmentation compared to the reference GTV contours by experts, with averaged Dice of 0.75 and 0.72 in training and testing sets respectively. By comparison, the two threshold-based approaches led to completely inappropriate delineation (averaged Dice, 0.114–0.311). Indeed, because it was applied automatically without specific constraints in the bounding box, it often included brain and node tissues (examples in Fig. 6a). Prognostic performances of models obtained using automatic segmentations with/without harmonization are detailed in Supplementary E5 and Tables S7, S8, and S9. The best result for each segmentation is presented in Table 5. Boxplots of Dice and *C*-index are displayed in Fig. 6b and c. Surprisingly, despite relying on inappropriate volumes of interest (extracting features from non-tumoral tissues in addition to the tumor volume), the models based on 41% SUV_{max} demonstrated better performance with a *C*-index of 0.704 achieved. Using the SUV above 2.5 segmentations, the *C*-index was

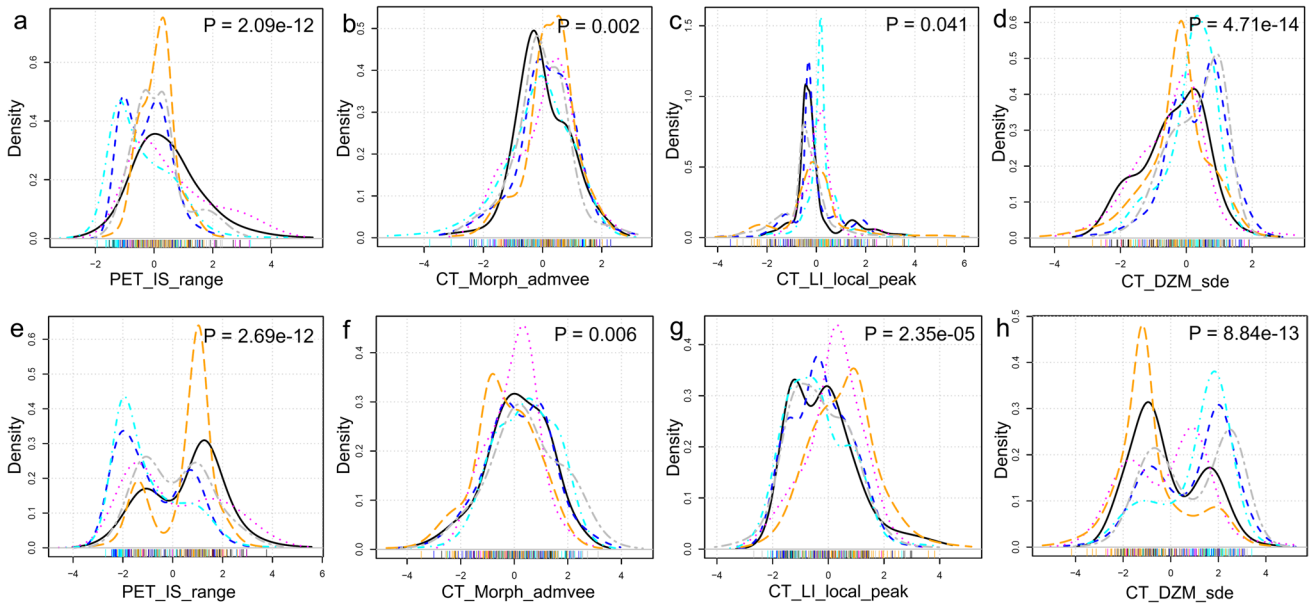

Fig. 4 The distribution of four selected features among six centers before (**a**–**d**) and after (**e**–**h**) ComBat harmonization. After ComBat strategy of AGK, the "batch-effect" was decreased for three features (decreased *p*-value of one-way ANOVA test), except one feature (CT_LI_local_peak). ComBat strategy of AAC reduced batch-effect for all four features, which is provided in Fig. S2

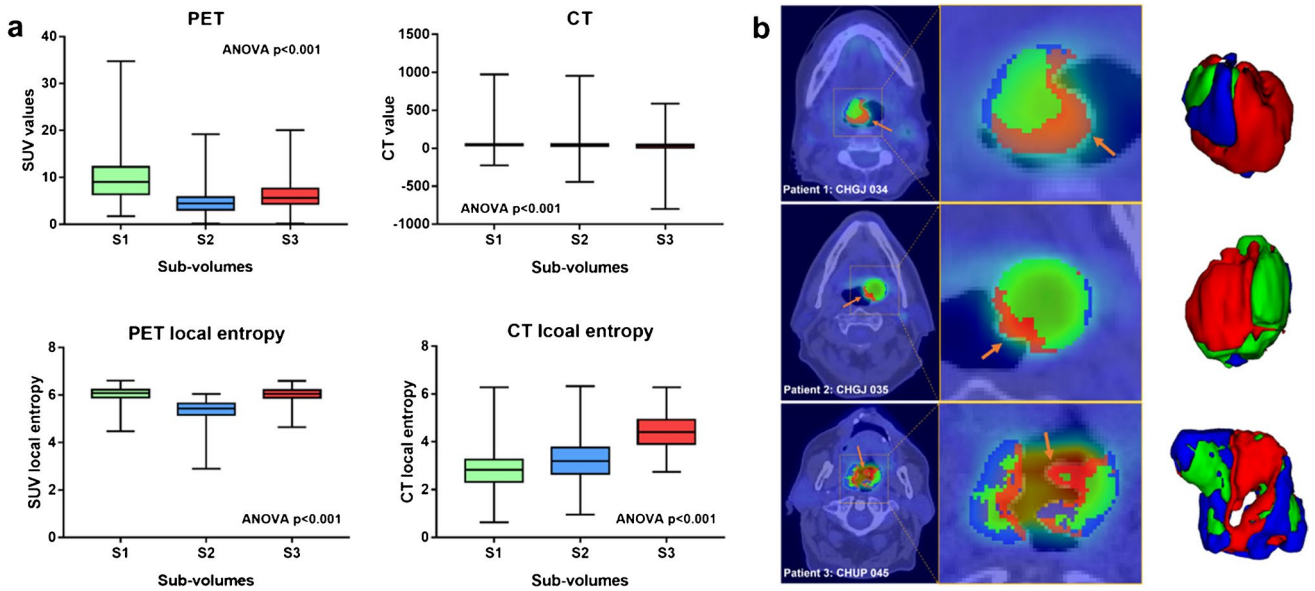

Fig. 5 Four imaging inputs in each of three sub-volumes in the training set (**a**). Results of intra-tumoral partitioning of three patients (**b**). Sub-volumes 1, 2, and 3 are indicated by green, blue, and red, respectively

0.685. Finally, the U-Net segmentation, despite providing a good overlap with the expert contours, led to a *C*-index of 0.674. Although the performance of models through automatic segmentations did not outperform the manual GTV (*C*-index 0.719), they still demonstrated their potential in prognosis prediction.

Comparison with other challengers' results

Table 5 lists the best results obtained in the HECKTOR 2021 challenge by other participants, relying on either a standard radiomics machine learning (ML) approach or deep learning (DL) models. Our best model (*C*-index, 0.719) reached

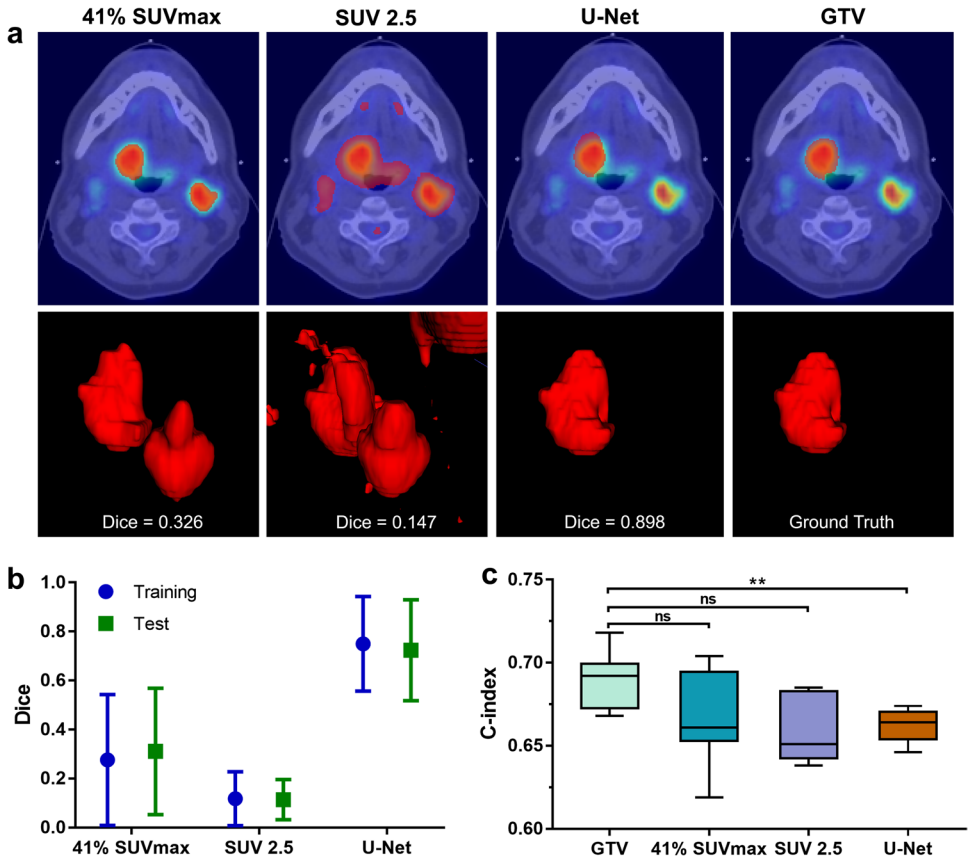

Fig. 6 Results of three automatic segmentations (Patient ID: CHGJ-018), with the Dice of 0.326, 0.147, and 0.898 respectively (**a**). Dice distributions of the training and testing sets (**b**). Boxplot of *C*-index values with and without ComBat strategies in the context of automatic segmentations and the manual GTV (**c**). Significance was calculated by the paired non-parametric *t*-test (ns, not statistically significant, $^{**}p < 0.01$)

Table 5 Comparison with the results of automatic segmentation and HECKTOR challenge on the same dataset

	Model	Clinical features	Harmonization	GT	Cohort	Dice	*C*-index	*p*-value	IBS
Manual GTV	ML	Yes	Yes	Yes	Training	–	0.698	1.11e − 05	0.146
					Testing	–	0.719	7.20e − 04	0.193
41% SUVmax	ML	Yes	Yes	No	Training	0.276	0.750	5.65e − 06	0.133
					Testing	0.311	0.704	0.022	0.193
SUV 2.5	ML	Yes	Yes	No	Training	0.118	0.681	1.22e − 04	0.144
					Testing	0.114	0.685	0.002	0.201
U-Net	ML	Yes	Yes	No	Training	0.750	0.693	1.25e − 04	0.146
					Testing	0.720	0.674	0.014	0.205
Saeed et al. [30]	DL	Yes	No	No (Task 2)	Testing	–	0.720	–	–
Naser et al. [32]	DL	Yes	No	Yes (Task 3)	Testing	–	0.698	–	–
Salmanpour et al. [28]	ML	Yes	No	No (Task 2)	Testing	0.680	0.683	–	–
Murugesan et al. [31]	DL	Yes	No	No (Task 2)	Testing	0.780	0.660	–	–
Starke et al. [29]	ML	Yes	No	Yes (Task 3)	Testing	–	0.659	–	–

Abbreviation: GT, ground truth segmentation that was manual delineated by clinicians; *Dice*, the mean values across the training or testing set; *p-value*, log-rank test; *IBS*, integrated Brier score

superior performance compared to the ML approaches similar to ours (0.683 [28] and 0.659 [29]), whereas it was found close to the best result (*C*-index, 0.720 [30]) and better than the other DL approaches (0.660 [31] and 0.698 [32]).

Discussion

The present study investigated the use of a radiomics pipeline applied to PET/CT images combined with different ComBat strategies to build prognostic models in a multicentric HNC cohort. Our results suggest that the standardized radiomics features without any harmonization showed relatively good robustness with respect to multicentric variations, providing models with predictive power slightly higher than relying on clinical factors alone. However, some of the investigated harmonization strategies could further improve the predictive ability of the resulting models. Although this improvement was moderate, it would still be enough to win the challenge (Task 3) or at least be among the best results (Table 5). We also evaluated a previously proposed approach to characterize intra-tumoral spatial heterogeneity and the resulting sub-volume signature could provide complementary prognostic information beyond the established clinical-radiomics model, and consistently improve the accuracy of risk stratification. Finally, automatic segmentations were applied to evaluate the feasibility of a fully automated radiomics pipeline, which exhibited close but slightly lower performance than relied on manually contours.

In this work, we only used IBSI standardized features [23] without any complex filtering. Deep features and more sophisticated modeling algorithms were also not considered, such as DeepCox [33] and random survival forest [34]. They are promising, however are more time-consuming and potentially increase the burden of parameter optimization and model interpretation, especially given the fact we had numerous models to compare. Our modeling process we chose to implement is computationally efficient and relatively stable. When randomly selecting other validation sets (20% of the training set), the same four features were finally selected, which suggests a good stability. We emphasize that simplicity is potentially an important factor to build reliable predictive models as well as to improve the generalization ability of the trained model. This point echoes the opinion in the overview of the 2021 HECKTOR challenge [22]. Previous studies further compared the performance of PET and CT separately [35], and showed equally good discriminative power for the two modalities with complementary prognostic information. Our results are also in line with this observation (see Supplementary Table S10 for results using only CT or PET), and we only reported resulting combining features from the two modalities since it led to the best performance.

This study is a relatively comprehensive investigation for the potential impact of ComBat harmonization, which had not been specifically addressed by most of the participants during the challenge [22]. The key point we observed is that ComBat harmonization led to either degraded or improved performance, depending on the chosen strategy. This suggests it should be performed separately for specific tumor type or patient population, but also separately for specific features or feature groups. In this study, we first detected multicentric effect in features (i.e., ANOVA test) and then split them into different groups for ComBat separately. Our results suggest that applying ComBat to feature groups separately improves performance compared to harmonizing all features together. In the latest guideline on the use of ComBat [16], it was stated that the ComBat transformation should be applied to the data affected by an imaging effect in the same way. Moreover, we found that applying ComBat before FS did not yield satisfactory results (Table 4), which is consistent with a recent study [36]. One hypothesis is that the features affected by heterogeneous imaging effects were put together to determine an identical transformation, so an adjustment appropriate for each feature cannot easily be determined. In our study, only four features were identified after FS; thus, we could easily observe their distributions separately and then adjust them by feature-specific transformation. That may partially explain why previous studies applying ComBat directly failed to obtain better results [16]. Both of ComBat decisions (i.e., for feature groups separately and after FS) indicated that applying ComBat in a relatively small feature group seems to be more effective. When taking this to the extreme, i.e., each feature is its own group, this basically comes down to *z*-standardization per center that was explored by previous studies [37, 38]. We also compared this strategy with our methods (Supplementary Material E6). It was shown that the performance of *z*-standardization per center (testing *C*-index, 0.587, 0.672, and 0.687 for radiomics, CR, and CRS model, respectively) was somewhere between the performance of the ComBat applied before FS and after FS, and between all features and feature groups (Table 4). This suggests that the ComBat strategy exploiting shared information between features is beneficial to reduce the batch-effect while maintain the biological variability of the dataset. Furthermore, given the imaging effect exists between and within centers, the labels determined by unsupervised clustering algorithms (i.e., K-means, hierarchical clustering) showed potential, which was also observed by a previous study where the imaging properties were highly heterogeneous [39]. In addition, the quantitative contributions of our ComBat strategies were assessed. The results showed that each choice (i.e., applying ComBat after FS, for feature groups and with clustering-determined labels) improved the *C*-index by 8.9%, 10.9%, and 4.4% compared to its alternative, respectively

(Supplementary Material E7). Both parametric and non-parametric ComBat [13] were tested, but we only report the results of the non-parametric version since it produced the best results.

Fully automated segmentations were implemented to investigate their potential impact in PFS prediction. A satisfying finding was that the segmentations generated through a U-Net model led to a slightly decreased but close performance compared to the reference manual contours, despite overlap coefficients below 0.8. A recent study explored the benefit of cleaning the contours specifically for radiomics in the same dataset, and reported that using dedicated contours performed the best in prediction [40], which is consistent with our results. However, a surprising finding was that completely inappropriate contours obtained through threshold approaches leading to include in the analysis non-tumoral areas (lymph nodes, but also brain or other physiological uptakes in the bounding box) led to models with good performance, in some cases even better than the ones focusing on the tumor only. This suggests that prognosis relevant information could be detected by automatic segmentation to a certain degree, which may exist not only in primary tumor but also in other metabolically active regions (i.e., lymph nodes). Moreover, several studies in the HECKTOR challenge reported simple segmentation methods, such as threshold and bounding box, outperforming the models obtained when using the provided reference contours (Task 3), either within the context of standard radiomics approaches or deep learning frameworks [22]. Sepehri et al. had similar findings in a lung cancer–related study [41]. These results potentially emphasize on the "non-essential" nature of dedicated tumor contours or overall delineation accuracy prior to extracting features. Nevertheless, our results based on the U-Net segmentation show the entire process can be fully automated, allowing for reproducible and large-scale radiomics studies.

Another objective of this study was to explore the intra-tumoral spatial heterogeneity characterization at subregional level through tumor volume partitioning. This step utilized patient information derived from both individual and population levels, and incorporated the multi-modality information [19, 20]. The volume of each sub-volume and their mutual relationships also with peri-tumor tissue were considered to develop an imaging signature. As shown in Fig. 5, it is an interesting finding that sub-volume 3 (red label) with high PET and CT entropy values is usually located in the border of the tumor, which may indicate the invasive border and expansive direction of tumor. Also, we found that tumors in the high-risk group were associated with a larger volume of sub-volume 3 and a larger size of its interaction with the tumor border (Supplementary Fig. S4). This finding is in line with the well-established biology research related to tumor aggressiveness [42]. Our prognostic analysis indicated that the sub-volume imaging signature could provide complementary information beyond the classical clinical-radiomics model. Especially for models without adequate predictive power of risk prediction, the sub-volume signature could help them to achieve significant risk stratification (Table 4).

Our study also presents some limitations. First, a significant difference regarding PFS was highlighted between the training and testing sets, which is a specific challenge of HECKTOR 2021. We therefore carefully checked the conditions of batch normalization, and plotted the calibration curves of our models (see Supplementary Material E8). It should be emphasized on that most of the models trained by us and other challengers did not exhibit a very large drop of performance between training and testing. The dataset has ~ 30% of patients with progression, which is consistent with the real-world population after radiotherapy treatment. Second, when constructing clinical and radiomics models, several features without strict significance in univariate analysis were included into the model, since they were automatically selected by the Lasso-Cox algorithm. Third, we only included textural features computed with a fixed bin number discretization, which may have been suboptimal for some features [21]. Moreover, although the proposed sub-volume signature showed consistently improved performance in prognosis, this improvement was small and not significant. Finally, compared to the results of the 2021 HECKTOR challenge, our models never exceeded the first place of Task 2 [34] (*C*-index, 0.720) which was obtained with a deep learning model with no segmentation. However, the advantage of our model is the use of standard features within classical radiomics pipeline, which could provide a better explainability and interpretability than a black box deep learning technique. In future work, we intend to replicate the present investigations in the larger datasets of the next HECKTOR editions.

Conclusion

The classical radiomics pipeline combined with the specific ComBat strategies was beneficial to predict PFS in a multicenter HNC cohort using pre-treatment PET/CT images. The intra-tumoral sub-volume characterization could provide complementary prognostic information beyond the established clinical-radiomics model. Furthermore, automatic segmentations embedding in radiomics pipeline exhibited the potential for prognosis that may obviate the need for dedicated tumor contours toward large-scale radiomics studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-023-06118-2.

Acknowledgements We thank the organizers of the HECKTOR 2021 challenge for authorizing the use of the dataset.

Author contribution Hui Xu, Nassib Abdallah, Jean-Marie Marion, Pierre Chauvet, Clovis Tauber, and Thomas Carlier searched relevant literatures and collected the data. Hui Xu, Lijun Lu, and Mathieu Hatt designed this study. Hui Xu, Nassib Abdallah, and Mathieu Hatt performed the data analysis and interpretation. Hui Xu and Mathieu Hatt drafted the primary manuscript, and all authors edited and reviewed it.

Funding This work was partly funded by (1) regions Bretagne, Pays de la Loire et Centre through the project HARMONY of the Cancéropole Grand Ouest; and (2) the National Natural Science Foundation of China under grants 81871437 and 12026601, and the China Scholarship Council under grant 202108440348.

Data availability Datasets are available through the challenge website of https://www.aicrowd.com/challenges/miccai-2021-hecktor.

Code availability Codes are available from the corresponding author on reasonable request.

Declarations

Ethics approval This is a retrospective study of a publicly available dataset. The requirement of informed consent was waived.

Conflict of interest The authors declare no competing interests.

References

1. Chow L. Head and neck cancer. N Engl J Med. 2020. https://doi.org/10.1056/NEJMra1715715.
2. Hatt M, Majdoub M, Vallieres M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J Nucl Med. 2015. https://doi.org/10.2967/jnumed.114.144055.
3. Lambin P, Leijenaar R, Deist TM, Peerlings J, de Jong E, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017. https://doi.org/10.1038/nrclinonc.2017.141.
4. Hatt M, Krizsan AK, Rahmim A, Bradshaw TJ, Costa PF, Forgacs A, et al. Joint EANM/SNMMI guideline on radiomics in nuclear medicine. Eur J Nucl Med Mol I. 2022. https://doi.org/10.1007/s00259-022-06001-6.
5. Hatt M, Cheze LRC, Antonorsi N, Tixier F, Tankyevych O, Jaouen V, et al. Radiomics in PET/CT: current status and future AI-based evolutions. Semin Nucl Med. 2021. https://doi.org/10.1053/j.semnuclmed.2020.09.002.
6. Hatt M, Le Rest CC, Tixier F, Badic B, Schick U, Visvikis D. Radiomics: data are also images. J Nucl Med. 2019. https://doi.org/10.2967/jnumed.118.220582.
7. Papadimitroulas P, Brocki L, Chung NC, Marchadour W, Vermet F, Gaubert L, et al. Artificial intelligence: deep learning in oncological radiomics and challenges of interpretability and data harmonization. Physica Med. 2021. https://doi.org/10.1016/j.ejmp.2021.03.009.
8. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. J Nucl Med. 2015. https://doi.org/10.2967/jnumed.115.156927.
9. Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015; https://doi.org/10.1007/s00259-014-2961-x.
10. Pfaehler E, van Sluis J, Merema BB, van Ooijen P, Berendsen RC, van Velden FH, et al. Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-dimensionally printed phantom inserts. J Nucl Med. 2020. https://doi.org/10.2967/jnumed.119.229724.
11. Llera A, Huertas I, Mir P, Beckmann CF. Quantitative intensity harmonization of dopamine transporter SPECT images using gamma mixture models. Mol Imaging Biol. 2019. https://doi.org/10.1007/s11307-018-1217-8.
12. Marcadent S, Hofmeister J, Preti MG, Martin SP, Van De Ville D, Montet X. Generative adversarial networks improve the reproducibility and discriminative power of radiomic features. Radiol Artif Intell. 2020. https://doi.org/10.1148/ryai.20201990035.
13. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007. https://doi.org/10.1093/biostatistics/kxj037.
14. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology. 2019. https://doi.org/10.1148/radiol.2019182023.
15. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018. https://doi.org/10.2967/jnumed.117.199935.
16. Orlhac F, Eertink JJ, Cottereau AS, Zijlstra JM, Thieblemont C, Meignan M, et al. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. J Nucl Med. 2022. https://doi.org/10.2967/jnumed.121.262464.
17. Da-ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci Rep. 2020. https://doi.org/10.1038/s41598-020-66110-w.
18. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res. 2015. https://doi.org/10.1158/1078-0432.CCR-14-0990.
19. Wu J, Gensheimer MF, Zhang N, Guo M, Liang R, Zhang C, et al. Tumor subregion evolution-based imaging features to assess early response and predict prognosis in oropharyngeal cancer. J Nucl Med. 2020. https://doi.org/10.2967/jnumed.119.230037.
20. Xu H, Lv W, Feng H, Du D, Yuan Q, Wang Q, et al. Subregional radiomics analysis of PET/CT imaging with intratumor partitioning: application to prognosis for nasopharyngeal carcinoma. Mol Imaging Biol. 2020. https://doi.org/10.1007/s11307-019-01439-x.
21. Vallieres M, Zwanenburg A, Badic B, Cheze LRC, Visvikis D, Hatt M. Responsible radiomics research for faster clinical translation. J Nucl Med. 2018. https://doi.org/10.2967/jnumed.117.200501.
22. Andrearczyk V, Oreiller V, Boughdad S, Rest CCL, Elhalawani H, Jreige M, et al. Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images. In: 3D head and neck tumor segmentation in PET/CT challenge. Springer; 2021. pp. 1–37. https://doi.org/10.1007/978-3-030-98253-9_1.
23. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020. https://doi.org/10.1148/radiol.2020191145.
24. Iantsen A, Visvikis D, Hatt M. Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. In: 3D head and neck tumor segmentation in PET/CT challenge. Springer; 2020. pp. 37–43. https://doi.org/10.1007/978-3-030-67194-5_4.
25. Kapp AV, Tibshirani R. Are clusters found in one dataset present in another dataset? Biostatistics. 2007. https://doi.org/10.1093/biostatistics/kxj029.

26. Efron B, Hastie T. Computer age statistical inference. Cambridge University Press. 2016. https://doi.org/10.1017/CBO9781316576533.
27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Journal of British Surgery. 2015. https://doi.org/10.1136/bmj.g7594.
28. Salmanpour MR, Hajianfar G, Rezaeijo SM, Ghaemi M, Rahmim A. Advanced automatic segmentation of tumors and survival prediction in head and neck cancer. In: 3D head and neck tumor segmentation in PET/CT challenge. Springer; 2021. pp. 202–210. https://doi.org/10.1007/978-3-030-98253-9_19.
29. Starke S, Thalmeier D, Steinbach P, Piraud M. A hybrid radiomics approach to modeling progression-free survival in head and neck cancers. In: 3D head and neck tumor segmentation in PET/CT challenge. Springer; 2021. pp. 266–277. https://doi.org/10.1007/978-3-030-98253-9_25.
30. Saeed N, Majzoub RA, Sobirov I, Yaqub M. An ensemble approach for patient prognosis of head and neck tumor using multimodal data. In: 3D head and neck tumor segmentation in PET/CT challenge. Springer; 2021. pp. 278–286. https://doi.org/10.1007/978-3-030-98253-9_26.
31. Murugesan GK, Brunner E, McCrumb D, Kumar J, VanOss J, Moore S, et al. Head and neck primary tumor segmentation using deep neural networks and adaptive ensembling. In: 3D head and neck tumor segmentation in PET/CT challenge. Springer; 2021. pp. 224–235. https://doi.org/10.1007/978-3-030-98253-9_21.
32. Naser MA, Wahid KA, Mohamed AS, Abdelaal MA, He R, Dede C, et al. Progression free survival prediction for head and neck cancer using deep learning based on clinical and PET/CT imaging data. In: 3D head and neck tumor segmentation in PET/CT challenge. Springer; 2021. pp. 287–299. https://doi.org/10.1007/978-3-030-98253-9_27.
33. Nagpal C, Yadlowsky S, Rostamzadeh N, Heller K. Deep Cox mixtures for survival regression. In: Machine learning for healthcare conference. PMLR; 2021. pp. 674–708. https://doi.org/10.48550/arXiv.2101.06536.
34. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008. https://doi.org/10.1214/08-AOAS169.
35. Bogowicz M, Riesterer O, Stark LS, Studer G, Unkelbach J, Guckenberger M, et al. Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. Acta Oncol. 2017. https://doi.org/10.1080/0284186X.2017.1346382.
36. Ferreira M, Lovinfosse P, Hermesse J, Decuypere M, Rousseau C, Lucia F, et al. [18F] FDG PET radiomics to predict disease-free survival in cervical cancer: a multi-scanner/center study with external validation. Eur J Nucl Med Mol I. 2021. https://doi.org/10.1007/s00259-021-05397-x.
37. Chatterjee A, Vallières M, Dohan A, Levesque IR, Ueno Y, Saif S, et al. Creating robust predictive radiomic models for data from independent institutions using normalization. IEEE Transactions on Radiation and Plasma Medical Sciences. 2019. https://doi.org/10.1109/TRPMS.2019.2893860.
38. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE. 2011. https://doi.org/10.1371/journal.pone.0017238.
39. Masson I, Da-ano R, Lucia F, Doré M, Castelli J, Goislard De Monsabert C, et al. Statistical harmonization can improve the development of a multicenter CT-based radiomic model predictive of nonresponse to induction chemotherapy in laryngeal cancers. Med Phys. 2021; https://doi.org/10.1002/mp.14948.
40. Fontaine P, Andrearczyk V, Oreiller V, Abler D, Castelli J, Acosta O, et al. Cleaning radiotherapy contours for radiomics studies, is it worth it? A head and neck cancer study. Clinical and Translational Radiation Oncology. 2022. https://doi.org/10.1016/j.ctro.2022.01.003.
41. Sepehri S, Tankyevych O, Iantsen A, Visvikis D, Hatt M, Le Rest CC. Accurate tumor delineation vs. rough volume of interest analysis for 18F-FDG PET/CT radiomics-based prognostic modeling in non-small cell lung cancer. Front Oncol. 2021. https://doi.org/10.3389/fonc.2021.726865.
42. Pietras K, Östman A. Hallmarks of cancer: interactions with the tumor stroma. Exp Cell Res. 2010. https://doi.org/10.1016/j.yexcr.2010.02.045.