scientific reports

Check for updates

OPEN

Identification of CT radiomic features robust to acquisition and segmentation variations for improved prediction of radiotherapy-treated lung cancer patient recurrence

Thomas Louis[1,9]✉, François Lucia[1,2,3,9]✉, François Cousin[1], Carole Mievis[4], Nicolas Jansen[4], Bernard Duysinx[5], Romain Le Pennec[6,7], Dimitris Visvikis[3], Malik Nebbache[2], Martin Rehn[2], Mohamed Hamya[2], Margaux Geier[8], Pierre-Yves Salaun[6,7], Ulrike Schick[2,3], Mathieu Hatt[3], Philippe Coucke[4], Pierre Lovinfosse[1,9] & Roland Hustinx[1,9]

The primary objective of the present study was to identify a subset of radiomic features extracted from primary tumor imaged by computed tomography of early-stage non-small cell lung cancer patients, which remain unaffected by variations in segmentation quality and in computed tomography image acquisition protocol. The robustness of these features to segmentation variations was assessed by analyzing the correlation of feature values extracted from lesion volumes delineated by two annotators. The robustness to variations in acquisition protocol was evaluated by examining the correlation of features extracted from high-dose and low-dose computed tomography scans, both of which were acquired for each patient as part of the stereotactic body radiotherapy planning process. Among 106 radiomic features considered, 21 were identified as robust. An analysis including univariate and multivariate assessments was subsequently conducted to estimate the predictive performance of these robust features on the outcome of early-stage non-small cell lung cancer patients treated with stereotactic body radiation therapy. The univariate predictive analysis revealed that robust features demonstrated superior predictive potential compared to non-robust features. The multivariate analysis indicated that linear regression models built with robust features displayed greater generalization capabilities by outperforming other models in predicting the outcomes of an external validation dataset.

Medical imaging plays a key role in the diagnostic, staging and follow-up processes of early-stage non-small cell lung cancer (ES-NSCLC). The inclusion of image acquisitions in the planning and dose calculation steps of radiotherapy treatments enhance their effectiveness while limiting the dose exposition for the patient. Stereotactic body radiation therapy (SBRT) is the standard of care for inoperable ES-NSCLC[1,2]. Computed tomography (CT), thanks to its high geometrical accuracy, is routinely employed to perform reliable dose calculations for SBRT[3]. Fluorodeoxyglucose positron emission tomography CT ([^{18}F]FDG PET/CT) is a molecular imaging technique combining metabolic and anatomical evaluation. This dual approach enhances diagnostic accuracy, refines lung

[1]Division of Nuclear Medicine and Oncological Imaging, University Hospital of Liège, Liège, Belgium. [2]Radiation Oncology Department, University Hospital of Brest, Brest, France. [3]LaTIM, INSERM, UMR 1101, University of Brest, Brest, France. [4]Department of Radiotherapy Oncology, University Hospital of Liège, Liège, Belgium. [5]Division of Pulmonology, University Hospital of Liège, Liège, Belgium. [6]Nuclear Medicine Department, University Hospital of Brest, Brest, France. [7]GETBO INSERM UMR 1304, University of Brest, UBO, Brest, France. [8]Medical Oncology Department, University Hospital of Brest, Brest, France. [9]These authors contributed equally: Thomas Louis, François Lucia, Pierre Lovinfosse and Roland Hustinx. ✉email: Thomas.louis@chuliege.be; francois.lucia@chu-brest.fr

Figure 1. Full methodology chart. i. High dose and low dose chest computed tomography (CT) scans were conducted on patients of both centers following center's specific imaging protocols. In each scan, the lung lesion is delineated in 3D by the annotator A. The lung lesion of 50 randomly drawn center A patients were segmented by annotator B. ii. CT scans were rescaled to 1 × 1x1mm voxel size. CT scan intensity was discretized to a fix bin count of 64. Pyradiomics was used to extract 106 features. The radiomic feature values were submitted to a feature-specific transformation selected to fit a Gaussian distribution over all center A value. A Z-score normalization was conducted on the updated feature values. iii. Inter-annotator robustness of the radiomic features was assessed by calculating the Intraclass Correlation Coefficient (ICC) (3,1) between the values extracted from the segmentations of annotator A and annotator B in high dose and low dose CT scan. Inter-protocol robustness of radiomic features was assessed by calculating Lin's Concordance Correlation Coefficient (CCC) between the values extracted from the high dose CT scan and the low dose CT scan of each patient In both centers. Robust features were identified as features with all coefficient values higher than 0.75. iv. Univariate analysis was conducted on radiomic features divided into three groups: robust features, high-dose (HD) value of non-robust features, low-dose (LD) value of non-robust features. Individual feature ability to predict regional or distant recurrence, regional or distant recurrence at 3 years and regional or distant progression free survival was studied. Distribution of performances between groups were studied. Generalized linear models (GLM) and Cox Proportional-Hazards (CoxPH) models were developed using features from four groups: Robust features, low dose features, high dose features, all features. The trained models were used to attempt at predicting the three previously cited outcomes. Performance of the models originating from the different groups were compared. Figure abbreviations: Non-Small Cell Lung Cancer (NSCLC), Radiation therapy (RT), Stereotactic Body Radiation Therapy (SBRT), Volume of interest (VOI), (AI) Artificial Intelligence.

cancer staging and enables better treatment optimization and therapy response monitoring[4]. In the ES-NSCLC radiotherapy planning phase, [^{18}F]FDG PET/CT complements the planning CT, facilitating the accurate delineation of the target volume and the preservation of organs at risk[5].

The extraction of quantitative, high-dimensional information from medical images with data-characterization algorithms, is commonly referred to as "Radiomics"[6]. This process has been increasingly used in recent years to analyze and predict clinical outcomes and has been efficiently applied to study NSCLC[7–10]. Radiomics is considered less robust than other omics approaches because of the different processing steps undergone by an image from its acquisition to the feature extraction[11–13]. Additionally, the lack of standardization in clinical CT acquisition protocols introduces considerable variability, which emphasizes this lack of robustness[14,15]. Post-extraction harmonization techniques such as ComBat were developed to limit the variability of the data by correcting batch effect[16,17]. Such strategies nevertheless require meeting several specific assumptions to be efficient, which limits the application on real-world datasets[18]. Despite the efforts of organizations such as the Image Biomarker Standardisation Initiative (IBSI) to develop consistent radiomics analysis workflows and identify reliable features[19,20], it remains important to take one step further and come up with a robust approach specific to ES NCSLC treatment. This approach needs to be unaffected by the imaging protocol and segmentation and deliver unbiased information related to the patients rather than the process artifacts.

Numerous strategies have been considered to assess the robustness of radiomic procedures while minimizing patient radiation exposure from multiple CT imaging. Some studies were based on phantom models[21,22]. In vivo studies utilized several scans acquired at intentionally reduced dose levels[23] or modeling to generate images with different acquisition parameters from a single CT scan per patient[24–28]. While these studies mainly focused on understanding the influence of operational parameters on the radiomic features and on defining optimal imaging and analysis workflows[29], they may not represent the reality of the clinical situation, for which numerous criteria can differ between acquisitions. Additionally, many of these previous works suffer from using small dataset, which is a recurring limitation in radiomic studies.

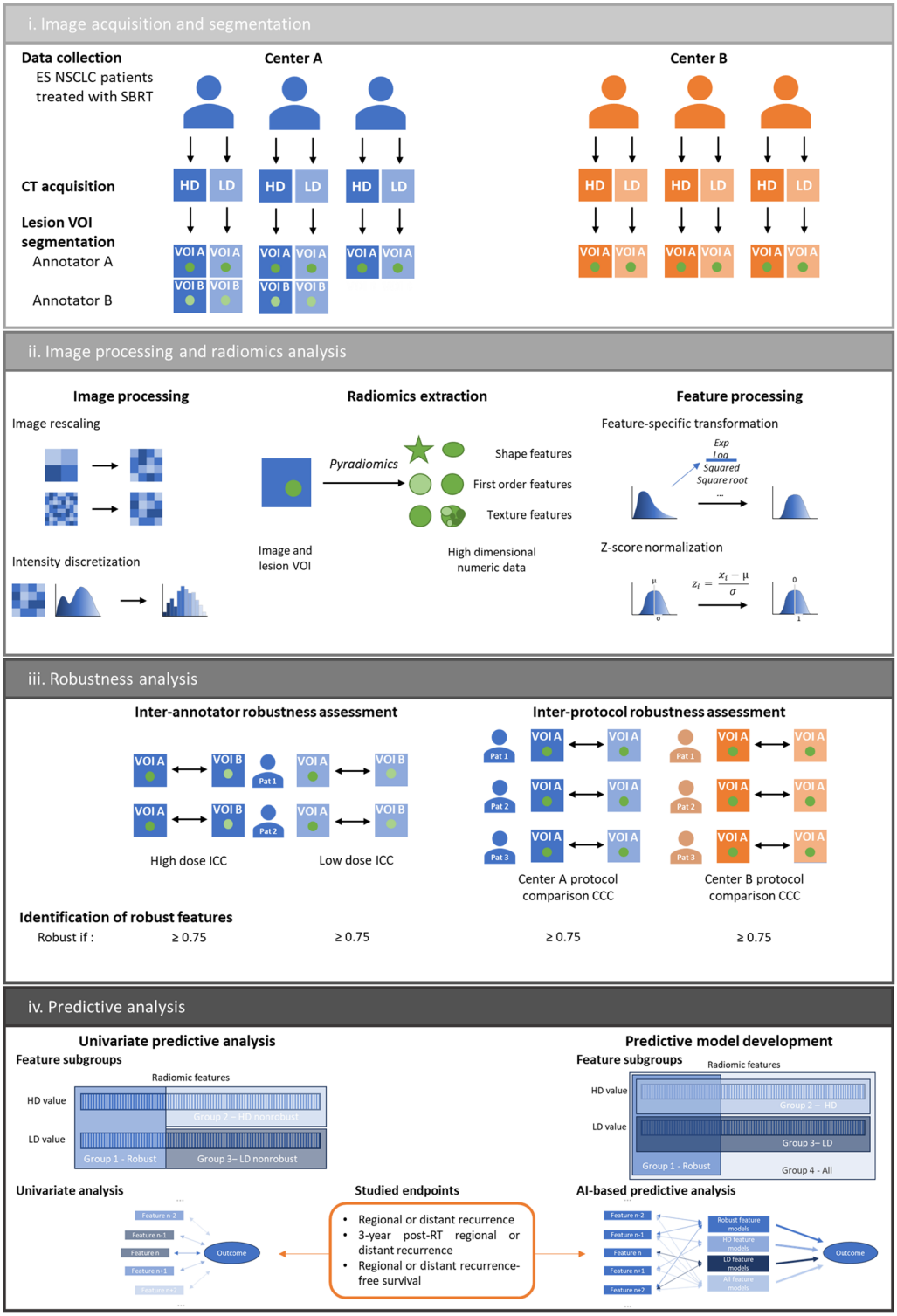
This study capitalizes on the planning protocol of radiotherapy for ES-NSCLC, which includes both high-dose dosimetric CT and low-dose CT of PET/CT scans for each patient. This specific scenario constitutes a great opportunity to provide a realistic insight into the effects of protocol modification and uncontrolled operating variations on radiomic feature robustness. The primary objective of this work was to transcend the influence of operational conditions and to identify features which are robust to variations in imaging protocols and lung cancer segmentations. Furthermore, a multicentric preliminary assessment of the predictive properties of these robust features was conducted. The analysis was focused on the occurrence of regional or distant relapse in patients treated with SBRT. Distant metastases are more frequent than local relapse[2] and harder to predict which justifies the interest of studying these endpoints[30,31].

Methods

The procedure workflow followed in this study is illustrated and described in Fig. 1.

Patients' selection

Patients with diagnosed ES-NSCLC T1 (< 3 cm) and T2 (3–5 cm) treated with lung SBRT between April 2010 and December 2020 were retrospectively collected in the databanks of two institutions: University Hospital (CHU) of Liège in Belgium—the center A—and University Hospital of Brest in France—the center B. This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of CHU of Liège (protocol code 2022/285; date of approval: 08 November 2022). The requirement of written informed consent from enrolled subjects was waived by the Institutional Review Board of CHU of Liège due to the retrospective study design.



Of the 597 eligible patients, patients were not included because of [^{18}F]FDG PET/CT was not available (n = 50), poor-quality of PET/CT or CT images (n = 57), or a delay between PET/CT and planning CT imaging longer than 8 weeks (n = 56). A total of 401 patients were eventually selected at center A and 33 at center B (see patient selection flowchart in Supplementary Fig. S1).

Endpoints

Regional recurrence was defined as lymph node metastasis in the bilateral hilar, mediastinal, or supraclavicular lymph node stations. Distant recurrence was defined as failure in the same pulmonary lobe (farther than 1.5 cm from the primary tumor), in other lung lobes (ipsi or contralateral lung) or in other organs. These recurrences had to be confirmed by histology or by a multidisciplinary committee on the basis of CT and [^{18}F]FDG PET/

CT. Recurrence was distinguished from second primary lung tumors by considering the pathology results, the interval between the recurrence and primary tumor, and the location of the recurrence in relation to the SBRT field[32]. Regional or distant recurrence (binary), recurrence after 3 years from the first day of RT (binary) and recurrence-free survival (survival) were considered as endpoints in this study.

Imaging modalities

High-dose (HD) CT scans of planning CT studies were captured with 2 types of scanners and 2 types of acquisition. In the CHU of Liège, expiration CT studies were acquired with a Philips Brilliance Big Bore CT (Philips Healthcare, Andover, MA, USA) and reconstructed with E kernel and a 1 mm slice thickness. In the CHU of Brest, free-breathing CT studies were performed with a Siemens Somatom (Siemens Healthcare, Malvern, PA, USA). No contrast-enhancing agent was used for the planning CT scan and reconstructed with filtered back projection B30f. kernel and a 2 mm slice thickness.

Low-dose (LD) CT scans from PET-CT studies were acquired with 3 types of scanners and 3 types of acquisition. In the CHU of Liège, studies were acquired using cross-calibrated Philips Gemini TF or BB (Philips Healthcare, Andover, MA, USA) and reconstructed with B kernel and a 3 mm slice thickness. In the CHU of Brest studies were performed with a Siemens Biograph mCT between 2016 and 2018 and with a Siemens digital Biograph Vision 600 between 2019 and 2020 (Siemens Healthcare, Malvern, PA, USA) and reconstructed using iterative reconstruction I30f kernel and a 2 mm slice thickness (see Supplementary Table S1 online for detailed acquisition information).

Volume segmentation

All pulmonary tumor were semi-automatically segmented on HD and LD CT scans following a previously validated semi-automatic approach exploiting an open-source software, *3D Slicer* (Slicer.org), and the Growcut algorithm by an experienced radiation oncologist (F.L.) —annotator A[33]. To evaluate the influence of the lesion volume of interest (VOI) variation on radiomic feature values, and hereby the inter-annotator robustness, the tumor of 50 randomly-selected patients was segmented by another expert (F.C.)—annotator B—using a semi-automatic approach on the open-source software ITK-SNAP (itksnap.org) (see Supplementary Fig. S2 for segmentation examples).

Image preprocessing

CT scans and related segmentations were resampled to $1 \times 1 \times 1$ mm^3 voxels using cubic spline interpolation and nearest neighbor interpolation respectively. CT scan voxel intensity was converted to Hounsfield units (HU).

Radiomic feature extraction

The open-source python package *PyRadiomics* v3.0.1 was used to extract radiomic features from the CT scans after an intensity discretization with a fixed bin count of 64[29,34]. Pyradiomics enabled the extraction of a total of 106 features from a 3D segmentation in the original non-derived image (see detailed feature list given in Supplementary Table S2 online). For each patient, 106 features were extracted from the lesion volume of the high-dose CT scan and 106 features from the lesion volume of the low-dose CT scan for a total of 212 features.

Robustness and predictive analysis

All the subsequent analyses were performed using R (v4.2.2) through RStudio (v 2023.06.1) IDE.

Inter-annotator robustness assessment

Dice Similarity Coefficient was used to compare segmentation between annotators[35]. The intraclass correlation coefficient (3,1) (ICC) was evaluated for each feature between annotator A and annotator B segmentation values[36]. Features with an ICC higher than 0.75 were considered as inter-annotator robust[37,38]. To express the influence of the patient's segmentation Dice score on the variation of extracted radiomic feature value, Z-score normalization was applied on each feature and the absolute variation of Z-score between radiomic values extracted from annotator A segmentation and from annotator B segmentation was calculated.

Inter-protocol robustness assessment

Lin's concordance correlation coefficient (CCC) was calculated for each feature between HD CT scan and LD CT scan values[39]. Features with an CCC higher than 0.75 were considered as inter-protocol robust following Altman's approach.

Radiomic feature processing

The features were submitted to a transformation specific to the feature distribution. First, it was assessed if the feature value distribution across all center A patients followed a gaussian distribution using Shapiro's test[40]. If not, it was assessed if the feature value distribution across all center A patients subjected to one of the following transformations followed a gaussian distribution: exponential, logarithmic, squared, cubed, square root, cubic root. If one of the transformed value distributions is significant to the Shapiro's test, the transformation is applied to the feature values of all center A and center B patients (see the list of transformation specific to each feature in Supplementary Table S3 online). All radiomic features were normalized using Z-score normalization. ComBat harmonization method was used on all radiomic features of center A and center B patients with center A patients considered as the reference set[16].

Univariate predictive analysis

Radiomic features were divided into three groups: the robust ones which featured an inter-annotator ICC and an inter-protocol CCC higher than 0.75, the HD feature value of the non-robust ones and the LD feature value of the non-robust ones. The ability of radiomic features to individually describe the regional or distant recurrence, regional or distant recurrence at 3 years and the regional or distant recurrence-free survival endpoints was evaluated. To do so, for each feature, one univariate prediction model was trained and tested on the whole center A set and one on the whole center B to predict endpoints. Generalized linear models (GLM) were used for binary endpoints and Cox progressional-hazards (Cox PH) model for the survival endpoint. The areas under the curve (AUC) of the ROC curve were calculated as performance metrics for binary endpoints and concordance for the survival endpoint. The oriented odds ratio (OR) of standardized radiomic feature for binary endpoints and hazard ratio (HR) for the survival endpoint were evaluated for each center by applying Z-score normalization over the whole set and calculating the OR/HR or the inverse of the OR/HR if it was less than one to ensure a greater-than-one value. The distribution of AUC/concordance and OR/HR between the different groups of radiomic features were then compared using Wilcoxon–Mann–Whitney tests with Bonferroni–Holm correction[41,42].

Multivariate predictive analysis

The data available for each patient was distributed into four categories: Robust radiomic features, all radiomic features extracted from the HD CT scan, all radiomic features extracted from the LD CT scan, both HD and LD CT scan radiomic features. Center A dataset was first stratified in a train set and an internal validation set with a 70/30 split. A fivefold cross validation approach was used on the train set for the training and signature selection steps. For each fold, the training phase started with a feature selection using the redundancy maximum relevance (mRMR) (F-test correlation quotient (FCQ) variation for binary endpoints and Wald-test correlation quotient (WCQ) variation for the survival endpoints) on the train subset and keeping the 10 first selected features for the next step[43]. When studying binary endpoints, correction for unbalanced data was conducted on the train subset using a combination of Synthetic Minority Over-sampling Technique (SMOTE) oversampling and Tomek links undersampling[44,45]. GLM models (resp. CoxPH models) with all possible combinations of the 10 selected features as signatures were trained on the train subset to predict binary (resp. survival) endpoints. The chosen predictive model for a specific outcome and a subset of features was selected among all signatures using the one standard error rule based on the Akaike information criterion (AIC) over the fivefold cross[46,47]. The performance metrics (AUC for binary endpoints and concordance for survival endpoint) of the trained model were evaluated on the whole train set, the internal validation set and the center B dataset used as external validation set. The whole process was repeated 10 times to evaluate the stability of the signatures selected and the predictive results. The distribution of AUC of the models built to a predict binary endpoint over the 10 reps and the distribution of concordance for the survival models over 10 reps were compared between the four feature subsets using Wilcoxon-Mann–Whitney tests with Holm correction.

Results

Among 434 selected patients who underwent SBRT for ES-NSCLC, regional and distant recurrence were found in 72 (17%) and 113 patients (26%), respectively, without significant differences between cohorts (see Supplementary Table S4 for full patient characteristics).

Inter-annotator robustness

The comparison of the segmentations from both annotators of the lung tumors of 50 randomly drawn patients resulted in a median Dice similarity coefficient of 0.74 (Q1–Q3: 0.66–0.83) for HD scans and 0.73 (Q1–Q3: 0.66–0.79) for LD scans. ICC (3,1) was calculated between feature values from annotator A's and annotator B's segmentations (given in Table 1). Out of the 106 extracted features, 59 (56%) had an ICC (3,1) greater than or equal to 0.75 for HD scans; 48 out of 106 (45%) had an ICC (3,1) greater than or equal to 0.75 for LD scans; 40 out of 106 (38%) had an ICC (3,1) greater than or equal to 0.75 for both acquisition protocols (Fig. 2.a). With a median value of 0.80 for HD CT scans (Q1–Q3: 0.53–0.89) and 0.71 for LD CT scans (Q1–Q3: 0.44–0.84), radiomic features exhibited an overall higher robustness in HD CT scans. The inter-annotator robust features were composed of shape and texture features, while no first-order intensity feature was selected. As shown in Fig. 2.b and 2.c presenting the distribution of absolute variation of radiomic feature Z-score for each patient in function of his/her segmentation Dice Score, the robust features were less influenced by the segmentation similarity than the non-robust ones.

Inter-protocol robustness

Lin's CCC was calculated for each radiomic feature between HD and LD scan values in center A and center B (shown in Fig. 3a and given in Table 1). Out of the 106 extracted features, 35 (33%) had a CCC greater than or equal to 0.75 in center A; 21 out of 106 (20%) had a CCC greater than or equal to 0.75 in center B; 21 out of 106 (20%) had a CCC greater than or equal to 0.75 in both centers. With a median value of 0.68 (Q1–Q3: 0.56–0.81) in center A and 0.63 (Q1–Q3: 0.44–0.75) in center B, radiomic features exhibited an overall higher robustness in center A. The inter-protocol robust features were composed of shape and texture features, while no first-order intensity feature was selected. All inter-protocol robust features were also inter-annotator robust. As shown in Fig. 3b,c, the delay between the two scans, which was limited to 56 days, has little influence on the absolute variation of radiomic feature Z-score between the scans independently of the feature robustness status.

Radiomics features	ICC HD	ICC LD	CCC center A	CCC center B
shape_LeastAxisLength	**0.97**	**0.89**	**0.92**	**0.90**
shape_Maximum2DDiameterSlice	**0.89**	**0.90**	**0.89**	**0.87**
shape_MeshVolume	**0.93**	**0.86**	**0.92**	**0.76**
shape_MinorAxisLength	**0.90**	**0.91**	**0.93**	**0.87**
shape_SurfaceArea	**0.89**	**0.89**	**0.93**	**0.85**
shape_SurfaceVolumeRatio	**0.88**	**0.88**	**0.88**	**0.88**
shape_VoxelVolume	**0.93**	**0.86**	**0.92**	**0.76**
glcm_Id	**0.94**	**0.82**	**0.88**	**0.86**
glcm_Idm	**0.95**	**0.82**	**0.87**	**0.83**
glcm_Idn	**0.85**	**0.82**	**0.83**	**0.79**
glcm_InverseVariance	**0.91**	**0.79**	**0.87**	**0.85**
gldm_DependenceNonUniformityNormalized	**0.92**	**0.89**	**0.89**	**0.91**
gldm_GrayLevelNonUniformity	**0.97**	**0.97**	**0.89**	**0.97**
gldm_SmallDependenceEmphasis	**0.88**	**0.82**	**0.88**	**0.82**
glrlm_GrayLevelNonUniformity	**0.99**	**0.97**	**0.92**	**0.95**
glrlm_RunLengthNonUniformityNormalized	**0.95**	**0.84**	**0.87**	**0.83**
glrlm_RunPercentage	**0.96**	**0.86**	**0.85**	**0.78**
glrlm_ShortRunEmphasis	**0.95**	**0.82**	**0.86**	**0.83**
glszm_GrayLevelNonUniformity	**0.89**	**0.87**	**0.92**	**0.78**
glszm_ZonePercentage	**0.89**	**0.80**	**0.89**	**0.84**
ngtdm_Coarseness	**0.81**	**0.87**	**0.77**	**0.86**
shape_Elongation	0.53	0.64	0.44	0.37
shape_Flatness	0.78	0.64	0.57	0.29
shape_MajorAxisLength	0.84	0.85	0.88	0.60
shape_Maximum2DDiameterColumn	0.90	0.86	0.90	0.70
shape_Maximum2DDiameterRow	0.77	0.82	0.89	0.74
shape_Maximum3DDiameter	0.82	0.86	0.89	0.72
shape_Sphericity	0.41	0.61	0.47	0.68
firstorder_10Percentile	0.47	0.22	0.66	0.52
firstorder_90Percentile	0.83	0.74	0.47	0.16

(continued)

Univariate predictive analysis

The predictive performances of the features distributed into the three groups labeled robust, non-robust HD and non-robust LD were evaluated in center A and in center B. The AUC and the oriented OR were calculated for the regional or distant recurrence start endpoint. The comparison of distributions of AUC and oriented OR between the three groups in center A and center B are shown in Fig. 4a–d. Slightly higher AUC and oriented OR were observed in robust features compared to HD and LD non-robust ones in centers A and B.

Univariate results for the regional or distant recurrence at 3 years post-RT endpoint are shown in Fig. 4e–h. Significantly higher AUC and oriented OR were observed in robust features compared to HD and LD non-robust ones in centers A and B.

The univariate analysis of the feature ability to predict regional or distant recurrence free survival gave analogous results (shown in Fig. 4i–l). Significantly higher concordance and higher oriented HR were observed in robust features compared to HD and LD non-robust ones in centers A and B.

See Supplementary Data S1 for a comprehensive statistical description of the univariate analysis results. See Supplementary Table S5a–c for detailed univariate analysis results on regional or distant recurrence prediction, 3-year post-RT regional or distant recurrence prediction and regional or distant recurrence free survival respectively.

firstorder_Energy	0.16	0.07	0.78	0.27
firstorder_Entropy	0.87	0.54	0.73	0.71
firstorder_InterquartileRange	0.39	-0.06	0.50	0.36
firstorder_Kurtosis	0.70	0.01	0.56	0.65
firstorder_Maximum	0.82	0.46	0.65	0.10
firstorder_MeanAbsoluteDeviation	0.25	-0.14	0.48	0.36
firstorder_Mean	0.67	0.68	0.67	0.34
firstorder_Median	0.64	0.74	0.70	0.35
firstorder_Minimum	0.33	0.01	0.67	0.49
firstorder_Range	0.47	0.09	0.64	0.36
firstorder_RobustMeanAbsoluteDeviation	0.33	-0.13	0.49	0.35
firstorder_RootMeanSquared	0.51	0.52	0.72	0.36
firstorder_Skewness	0.87	0.46	0.75	0.59
firstorder_TotalEnergy	0.16	0.07	0.78	0.27
firstorder_Uniformity	0.93	0.62	0.71	0.75
firstorder_Variance	0.12	-0.07	0.44	0.39
glcm_Autocorrelation	0.84	0.69	0.68	0.44
glcm_ClusterProminence	0.17	-0.12	0.45	0.37
glcm_ClusterShade	0.44	0.09	0.62	0.06
glcm_ClusterTendency	0.35	-0.01	0.51	0.45
glcm_Contrast	0.74	0.82	0.63	0.62
glcm_Correlation	0.58	0.60	0.34	0.30
glcm_DifferenceAverage	0.83	0.82	0.80	0.75
glcm_DifferenceEntropy	0.84	0.70	0.80	0.68
glcm_DifferenceVariance	0.60	0.76	0.38	0.42
glcm_Idmn	0.76	0.82	0.65	0.62
glcm_Imc1	0.57	0.64	0.68	0.49
glcm_Imc2	0.66	0.42	0.55	0.48
glcm_JointAverage	0.82	0.65	0.70	0.44
glcm_JointEnergy	0.89	0.71	0.64	0.71
glcm_JointEntropy	0.82	0.59	0.72	0.68
glcm_MCC	0.17	0.36	0.33	0.26
glcm_MaximumProbability	0.89	0.78	0.65	0.63
glcm_SumAverage	0.82	0.65	0.70	0.44
glcm_SumEntropy	0.83	0.54	0.67	0.63
glcm_SumSquares	0.44	0.03	0.56	0.53
gldm_DependenceEntropy	0.67	0.75	0.59	0.53
gldm_DependenceNonUniformity	0.77	0.77	0.82	0.60
gldm_DependenceVariance	0.89	0.85	0.72	0.55
gldm_GrayLevelVariance	0.43	0.04	0.57	0.54
gldm_HighGrayLevelEmphasis	0.85	0.69	0.70	0.48
gldm_LargeDependenceEmphasis	0.92	0.86	0.77	0.68
gldm_LargeDependenceHighGrayLevelEmphasis	0.90	0.91	0.80	0.66
gldm_LargeDependenceLowGrayLevelEmphasis	0.00	0.13	0.09	0.29
gldm_LowGrayLevelEmphasis	0.21	0.43	0.36	0.72

(continued)

gldm_SmallDependenceHighGrayLevelEmphasis	0.72	0.79	0.68	0.39
gldm_SmallDependenceLowGrayLevelEmphasis	0.70	0.61	0.59	0.74
glrlm_GrayLevelNonUniformityNormalized	0.89	0.44	0.71	0.75
glrlm_GrayLevelVariance	0.37	0.01	0.55	0.53
glrlm_HighGrayLevelRunEmphasis	0.84	0.64	0.68	0.49
glrlm_LongRunEmphasis	0.92	0.85	0.68	0.71
glrlm_LongRunHighGrayLevelEmphasis	0.91	0.86	0.77	0.65
glrlm_LongRunLowGrayLevelEmphasis	0.04	0.35	0.06	0.66
glrlm_LowGrayLevelRunEmphasis	0.39	0.43	0.72	0.72
glrlm_RunEntropy	0.57	0.12	0.54	0.53
glrlm_RunLengthNonUniformity	0.89	0.85	0.87	0.65
glrlm_RunVariance	0.89	0.85	0.58	0.63
glrlm_ShortRunHighGrayLevelEmphasis	0.80	0.57	0.65	0.43
glrlm_ShortRunLowGrayLevelEmphasis	0.52	0.44	0.68	0.74
glszm_GrayLevelNonUniformityNormalized	0.20	0.11	0.62	0.71
glszm_GrayLevelVariance	0.32	0.38	0.53	0.63
glszm_HighGrayLevelZoneEmphasis	0.70	0.41	0.56	0.48
glszm_LargeAreaEmphasis	0.96	0.97	0.72	0.96
glszm_LargeAreaHighGrayLevelEmphasis	0.95	0.95	0.74	0.94
glszm_LargeAreaLowGrayLevelEmphasis	0.42	0.96	0.09	0.94
glszm_LowGrayLevelZoneEmphasis	0.63	0.47	0.64	0.64
glszm_SizeZoneNonUniformity	0.86	0.83	0.87	0.70
glszm_SizeZoneNonUniformityNormalized	0.67	0.81	0.42	0.33
glszm_SmallAreaEmphasis	0.68	0.78	0.43	0.31
glszm_SmallAreaHighGrayLevelEmphasis	0.72	0.58	0.45	0.36
glszm_SmallAreaLowGrayLevelEmphasis	0.62	0.53	0.54	0.61
glszm_ZoneEntropy	0.60	0.72	0.36	0.32
glszm_ZoneVariance	0.96	0.97	0.72	0.96
ngtdm_Busyness	0.85	0.70	0.72	0.78
ngtdm_Complexity	0.59	0.81	0.51	0.48
ngtdm_Contrast	0.75	0.73	0.64	0.69

Table 1. Radiomic feature coefficient table. Complete list of intraclass correlation coefficient (ICC) and Lin's concordance correlation coefficient (CCC) for each feature. Robust features are highlighted in bold and placed at the beginning of the table.

The 21 features eventually selected were least axis length, maximum 2D diameter slice, mesh volume, minor axis length, surface area, surface volume ratio, voxel volume, Gray Level Co-occurrence Matrix (GLCM) inverse difference, GLCM inverse difference moment, GLCM inverse difference normalized, GLCM inverse variance, Gray Level Dependence Matrix (GLDM) dependence non-uniformity normalized, GLDM grey-level non-uniformity, GLDM small dependence emphasis, Gray Level Run Length Matrix (GLRLM) grey-level non uniformity, GLRLM run length non-uniformity normalized, GLRLM run percentage, GLRLM short run emphasis, Gray Level Size Zone Matrix (GLSZM) gray-level non-uniformity, GLSZM zone percentage and Neighboring Gray Tone Difference Matrix (NGTDM) coarseness.

Multivariate predictive analysis

For each subset of features and each endpoint, at each iteration of the five folds and of the ten repetitions, $2^{10} - 1$ signatures were generated with the 10 features selected by mRMR. The GLM and Cox PH models based on these signatures were trained and tested in the fivefold cross-validation and one signature was selected with the one

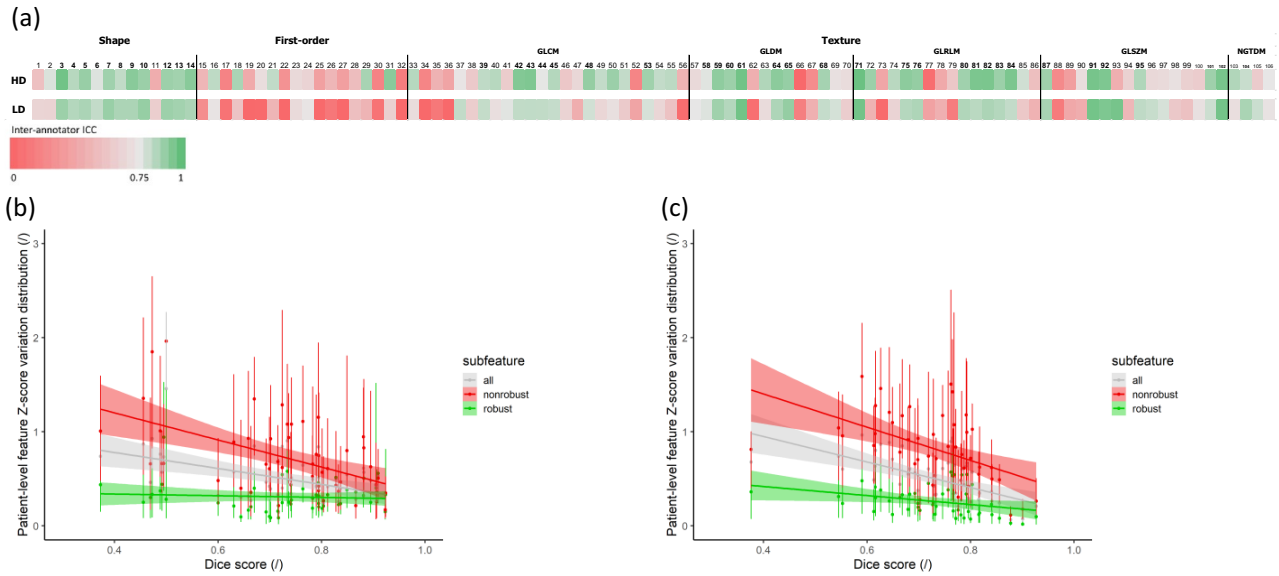


Figure 2. Radiomic feature inter-annotator robustness assessment. (**a**) Inter-annotator Intraclass Correlation Coefficient (ICC) of each radiomic feature in high dose and low dose CT scans. Inter-annotator robust features (highlighted in bold) were identified as features with both ICC values higher than 0.75. (**b**) Influence of the segmentation Dice score on feature Z-score variation in high dose scans. (**c**) Influence of the segmentation Dice score on feature Z-score variation in low dose scans.

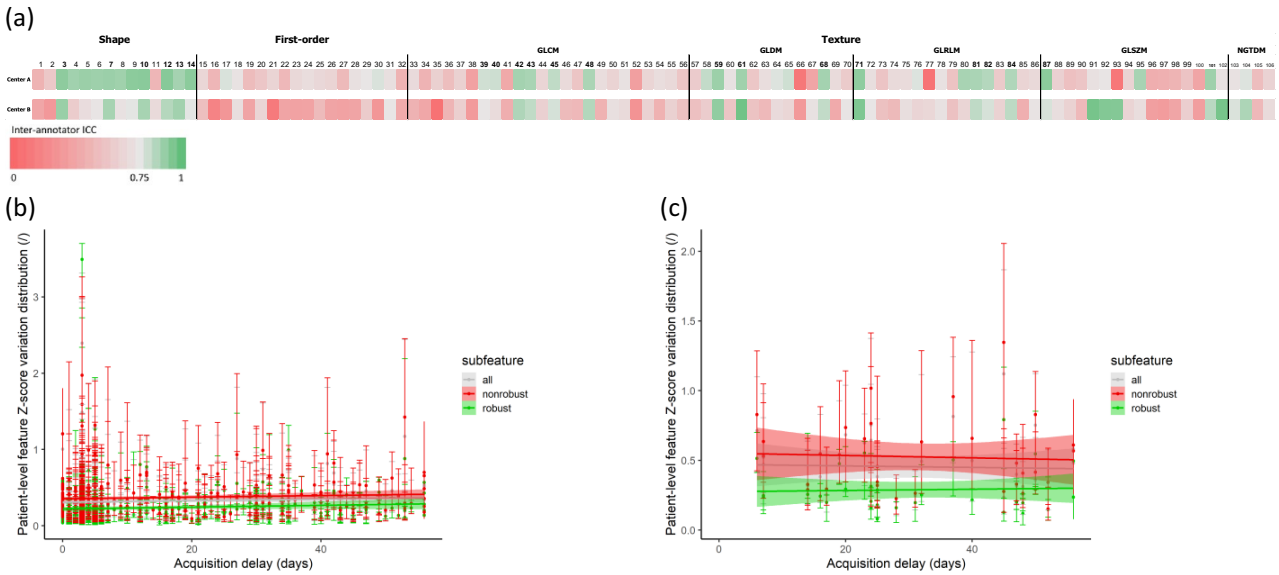


Figure 3. Radiomic feature inter-protocol robustness assessment. (**a**) Inter-protocol Lin's Concordance Correlation Coefficient (CCC) of each radiomic feature in patients from center A and from center B. Inter-protocol robust features (highlighted in bold) were identified as features with both CCC values higher than 0.75. (**b**) Influence of the delay between acquisitions on feature Z-score variation in center A. (**c**) Influence of the delay between acquisitions on feature Z-score variation in center B.

standard error rule based on the AIC. The distributions of AUC and concordance of the selected models used on the train, internal validation and external validation sets to predict regional or distant recurrence, 3-year post-RT regional or distant recurrence and regional or distant recurrence free survival are shown in Fig. 5a–c respectively. A comprehensive description of the multivariate analysis results is given in Supplemental Data S2.

Wilcoxon–Mann–Whitney tests with Holm correction were additionally performed to compare results on the train, internal validation and external validation sets separately (see full comparison results in Supplementary Table S6a–c online). The predictive performance of the models built with robust features on the train and internal validation set demonstrated similar-to-greater tendencies compared to other features groups. These robust

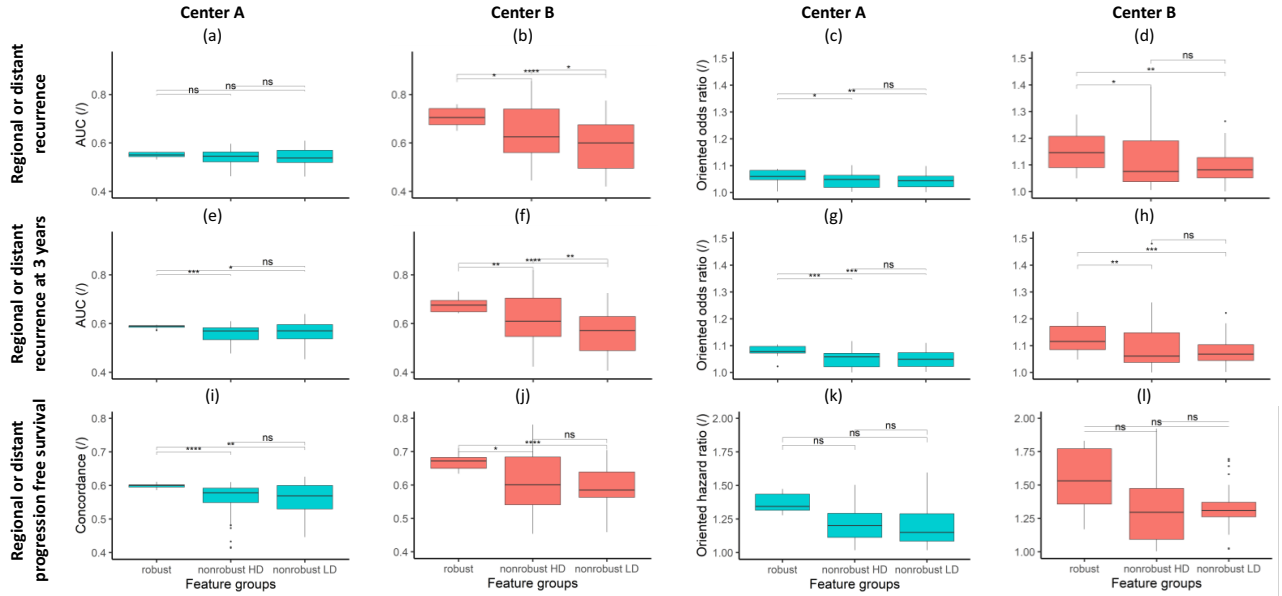


Figure 4. Univariate predictive analysis. Distributions of univariate AUC from the different feature groups for the prediction of regional or distant recurrence in center A (**a**) and center B (**b**), for the prediction of regional or distant recurrence at three years in center A (**e**) and center B (**f**). Distributions of odds ratio from the different feature groups for the prediction of regional or distant recurrence in center A (**c**) and center B (**d**), for the prediction of regional or distant recurrence at three years in center A (**g**) and center B (**h**). Distributions of univariate concordance from the different feature groups for the prediction of regional or distant progression free survival in center A (**i**) and center B (**j**). Distributions of hazard ratio from the different feature groups for the prediction of regional or distant progression free survival in center A (**k**) and center B (**l**).

feature-only models significantly outperformed the all-feature, HD-feature and LD-feature models in predicting the three considered endpoints in the external validation set.

See Supplementary Table S7a–c online for the individual performance of the model selected at each of the 10 repetitions for each feature subset to predict regional or distant recurrence prediction, 3-year post-RT regional or distant recurrence prediction and regional or distant recurrence free survival respectively.

Discussion

This study identified a subset of features invariant to imaging protocol variation and lesion segmentation quality. These features demonstrated a greater predictive potential for regional or distant recurrence in ES-NSCLC patients treated with SBRT through univariate and multivariate analyses.

Out of the 106 CT radiomic features studied, 40 (38%) had HD ICC(3,1) and LD ICC(3,1) values greater than or equal to 0.75 and were labeled as inter-annotator robust. The globally lower ICC in LD scans can be explained by a greater impact of segmentation variations on radiomic features extracted from lower resolution images. The features identified as robust does not seem to be influenced by the segmentation Dice score both in HD and LD scans even in poor cases (Dice score < 0.5). It means that these features are highly invariant to the segmentation quality and that the texture features should be less influenced by intensity values at the border of the segmentation. Shape features identified as robust between annotators were primarily determined by the size of the segmentations, whereas non-robust features were affected by their shapes. The diminished impact of the Dice score on robust shape features suggests that while the segmentations from the two annotators were similar in size, they must have exhibited slight variations in shape.

Features value variation analysis between HD and LD scans in center A and center B led to the selection of 21 (20%) inter-protocol robust features. The variation of value for both robust and non-robust features between HD and LD scans does not seem to be influenced by the delay between acquisition in the 8-week limit range. The 21 features selected as robust regarding the imaging protocol are also robust regarding the annotator.

Amongst the 21 features eventually selected, 7 are shape factors and the others are texture features, including 4 GLCM features, 3 GLDM features, 4 GLRLM, 2 GLSZM features and 1 NGTDM feature. The high proportion of shape features identified as robust (7 out of 14 shape features initially extracted) could be expected since the shape of the lesion should not be greatly influenced by the imaging protocol. No first-order intensity feature was selected as robust. This contradicted other studies which claimed that intensity features, along with shape features, are more robust than texture ones[48,49]. The high variability of first order statistics may nevertheless be explained by the significant difference in intensity between the tumor and its neighboring area leading to major intensity fluctuations at the border of the lesion. GLDM gray-level non-uniformity, GLSZM gray-level non-uniformity were already identified as reproducible in response to dose and kernel variations[24]. GLRLM gray level non-uniformity was reported as promising feature for NSCLC prognosis[50].

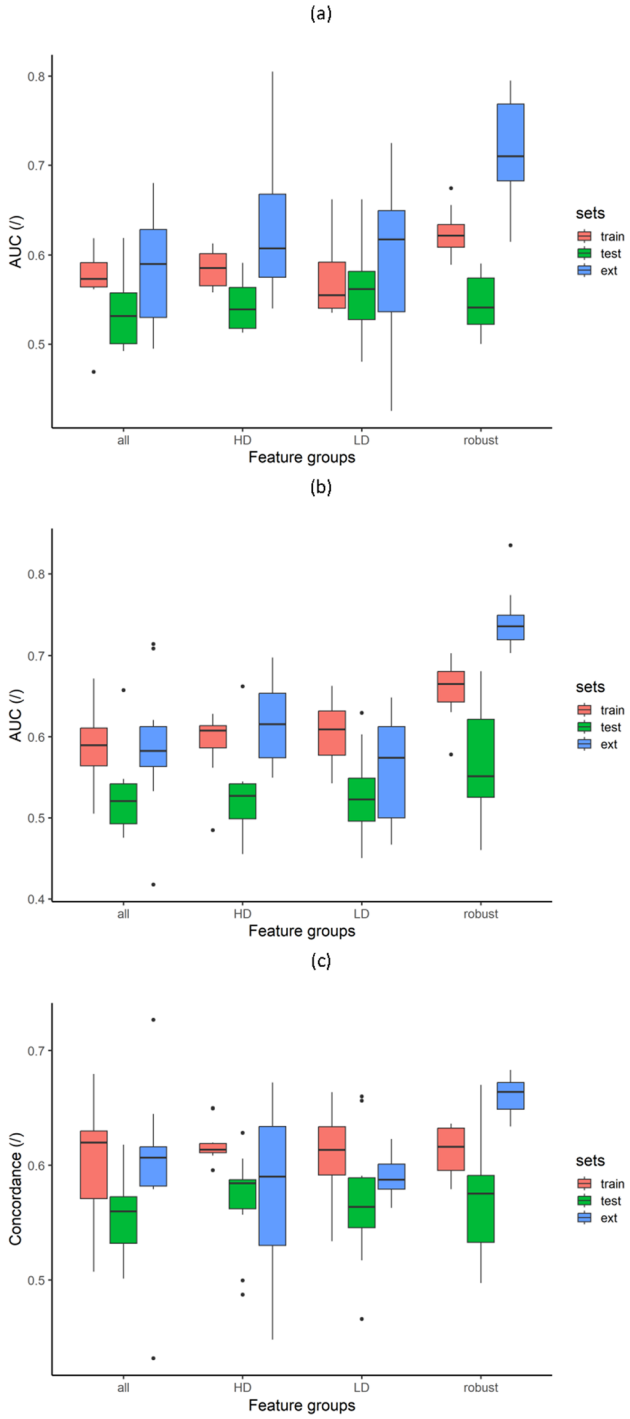


Figure 5. Multivariate predictive analysis. Prediction of region or distant recurrence (**a**), 3-year post-RT regional or distant recurrence (**b**) and regional or distant progression free survival (**c**) using generalized linear models (GLM), GLM and Cox Proportional-Hazards (CoxPH) models respectively.

The decision to consider the regional or distant recurrence after a fixed follow-up time of 3 years as an endpoint came from the high variability of follow-up time and time of occurrence between patients. A delay of 3 years after radiotherapy allowed to include most patients while featuring a sufficient level of occurrence.

Robust features exhibited higher AUC/concordance and oriented OR/HR than non-robust ones to predict regional or distant recurrence. This addresses the concern that, by only considering features with common information between protocols, we neglect the subtle but relevant information in the images[51,52]. Non-robust HD features exhibited better results than non-robust LD features indicating that higher dose images globally contain more information. While the median individual AUCs and concordance for the center A dataset were

between 0.55 and 0.6, indicating an overall limited univariate predictive power, this study preferably focused on the generalization of the selected features and the greater predictive potential which they demonstrated.

The multivariate predictive analysis resulted in similar tendencies for all endpoints. The similar-to-greater performances of the robust features-only models were consistent with what was observed in the univariate analysis. Even if the other feature groups also included the robust features in the multivariate analysis, the presence of the other features could dilute the information and the feature selection step before the model development was not sufficient to properly identify the relevant ones. The higher AUC/concordance of the robust feature-only model on the external validation set translated the better generalization ability of the robust radiomic features : while results for the train and test subsets, which both consisted of center A patients, were similar for signatures including robust and non-robust features, the greater performance of robust-feature only signatures on the external validation dataset composed of center B patients proved that non-robust features were not able to maintain their predictive potential when applied to datasets with other image acquisition conditions.

The present study differs from the literature by describing a multicentric approach which includes high-dose dosimetric CT and low-dose CT of PET/CT scans for each patient by benefiting from the planning protocol of radiotherapy for ES-NSCLC. It goes one step further than studies relying on phantom models[21,22] or artificially-perturbated scans[24–28], as it uses authentic unaltered patient scans with distinct acquisition protocols to identify both inter-annotator and inter-protocol robust radiomic features with a comprehensive and complete approach. The added value of the identified robust features is then brought forward by univariate and multivariate analyses.

It nevertheless suffers from some limitations. The dataset of CHU Brest is small and numerous patients were additionally excluded due to too much delay between high-dose and low-dose acquisition. The limited number of patients could have influenced the selection of features identified as robust in the inter-protocol robustness analysis. The CCC values were however globally lower for center B than center A resulting in a stricter feature selection due to center B. The higher predictive power of robust features in center B could also be interpreted with caution but similar relative tendencies between groups are observed for center A. While the feature selection could be generalized to other studies with ES-NSCLC patients, the results and discussions related to the univariate and multivariate predictive analysis are specific to patient response to ESBR treatment. Additional studies must be conducted to evaluate the generalization of these results to other clinical strategies.

To conclude, in this proof-of-concept study, we have identified a subgroup of features that are not affected by the segmentation quality and the imaging protocols. This group of features demonstrated greater predictive performance on outcomes of ES-NSCLC patients treated with SBRT. Limiting the model development to 21 features may overlook valuable information in more standardized acquisition protocols and more elaborate machine learning models. These features are nevertheless a solid basis for the development of models in multicentric studies.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 5 December 2023; Accepted: 1 April 2024
Published online: 19 April 2024

References

1. Postmus, P. E. *et al.* Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **28**, 1–21. https://doi.org/10.1093/ANNONC/MDX222 (2017).
2. Timmerman, R. *et al.* Stereotactic body radiation therapy for inoperable early stage lung cancer. *JAMA* **303**(11), 1070–1076. https://doi.org/10.1001/JAMA.2010.261 (2010).
3. Davis, A. T., Palmer, A. L. & Nisbet, A. Can CT scan protocols used for radiotherapy treatment planning be adjusted to optimize image quality and patient dose? A systematic review. *Br. J. Radiol.* **90**, 1076. https://doi.org/10.1259/BJR.20160406 (2017).
4. Vaz, S. C. *et al.* Joint EANM/SNMMI/ESTRO practice recommendations for the use of 2-[18F]FDG PET/CT external beam radiation treatment planning in lung cancer V1.0. *Eur. J. Nucl. Med. Mol. Imaging* **49**(4), 1386–1406. https://doi.org/10.1007/S00259-021-05624-5 (2022).
5. Gkika, E., Grosu, A. L. & Nestle, U. The use of 18F-FDG PET/CT for radiotherapy treatment planning in non-small cell lung cancer: A mini-review. *Precis. Cancer Med.* **6**, 1. https://doi.org/10.21037/PCM-22-38/COIF (2023).
6. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**(4), 441–446. https://doi.org/10.1016/J.EJCA.2011.11.036 (2012).
7. Mu, W. *et al.* Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat. Commun.* **11**(1), 1–11. https://doi.org/10.1038/s41467-020-19116-x (2020).
8. Sollini, M., Cozzi, L., Antunovic, L., Chiti, A. & Kirienko, M. PET Radiomics in NSCLC: State of the art and a proposal for harmonization of methodology. *Sci. Rep.* **7**(1), 1–15. https://doi.org/10.1038/s41598-017-00426-y (2017).
9. Frix, A. N. *et al.* Radiomics in lung diseases imaging: State-of-the-art for clinicians. *J. Pers. Med.* **11**(7), 602. https://doi.org/10.3390/JPM11070602 (2021).
10. Lovinfosse, P. *et al.* FDG PET/CT texture analysis for predicting the outcome of lung cancer treated by stereotactic body radiation therapy. *Eur. J. Nucl. Med. Mol. Imaging* **43**(8), 1453–1460. https://doi.org/10.1007/S00259-016-3314-8/FIGURES/2 (2016).
11. Park, J. E., Park, S. Y., Kim, H. J. & Kim, H. S. Reproducibility and generalizability in radiomics modeling: Possible strategies in radiologic and statistical perspectives. *Korean J. Radiol.* **20**(7), 1124. https://doi.org/10.3348/KJR.2018.0070 (2019).
12. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Phys. Med. Biol.* **61**(13), R150–R166. https://doi.org/10.1088/0031-9155/61/13/R150 (2016).
13. Limkin, E. J. *et al.* Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann. Oncol.* **28**(6), 1191–1206. https://doi.org/10.1093/ANNONC/MDX034 (2017).
14. Lee, S. H., Cho, H. H., Lee, H. Y. & Park, H. Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: A focus on lung cancer. *Cancer Imaging* **19**(1), 1–12. https://doi.org/10.1186/S40644-019-0239-Z/TABLES/5 (2019).
15. Mackin, D. *et al.* Measuring CT scanner variability of radiomics features HHS Public Access. *Invest. Radiol.* **50**(11), 757–765. https://doi.org/10.1097/RLI.0000000000000180 (2015).

16. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**(1), 118–127. https://doi.org/10.1093/BIOSTATISTICS/KXJ037 (2007).
17. Horng, H. *et al.* Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Sci. Rep.* **12**(1), 1–12. https://doi.org/10.1038/s41598-022-08412-9 (2022).
18. Orlhac, F. *et al.* A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J. Nucl. Med.* **63**(2), 172–179. https://doi.org/10.2967/JNUMED.121.262464 (2022).
19. Zwanenburg, A. *et al.* The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**(2), 328. https://doi.org/10.1148/RADIOL.2020191145 (2020).
20. Hatt, M. *et al.* Joint EANM/SNMMI guideline on radiomics in nuclear medicine. *Eur. J. Nucl. Med. Mol. Imaging* **50**(2), 352–375. https://doi.org/10.1007/S00259-022-06001-6 (2022).
21. Zhong, J. *et al.* Robustness of radiomics features of virtual unenhanced and virtual monoenergetic images in dual-energy CT among different imaging platforms and potential role of CT number variability. *Insights Imaging* **14**(1), 1–13. https://doi.org/10.1186/S13244-023-01426-5/TABLES/4 (2023).
22. Chen, Y. *et al.* Robustness of CT radiomics features: Consistency within and between single-energy CT and dual-energy CT. *Eur. Radiol.* **32**(8), 5480–5490. https://doi.org/10.1007/S00330-022-08628-3/TABLES/3 (2022).
23. Bartholomeus, G. A. *et al.* Robustness of pulmonary nodule radiomic features on computed tomography as a function of varying radiation dose levels—a multi-dose in vivo patient study. *Eur. Radiol.* **33**(10), 7044–7055. https://doi.org/10.1007/S00330-023-09643-8/FIGURES/6 (2023).
24. Emaminejad, N. *et al.* Reproducibility of lung nodule radiomic features: multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters. *Med. Phys.* **48**(6), 2906. https://doi.org/10.1002/MP.14830 (2021).
25. Bagher-Ebadian, H., Siddiqui, F., Liu, C., Movsas, B. & Chetty, I. J. On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. *Med. Phys.* **44**(5), 1755–1770. https://doi.org/10.1002/MP.12188 (2017).
26. Zhang, J. *et al.* Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on nasopharyngeal carcinoma patients. *Radiother. Oncol.* **183**, 1. https://doi.org/10.1016/j.radonc.2023.109578 (2023).
27. Teng, X. *et al.* Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma. *Front. Oncol.* **12**, 974467. https://doi.org/10.3389/FONC.2022.974467 (2022).
28. Teng, X. *et al.* Building reliable radiomic models using image perturbation. *Sci. Rep.* **12**(1), 1–10. https://doi.org/10.1038/s41598-022-14178-x (2022).
29. Escudero Sanchez, L. *et al.* Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle. *Sci. Rep.* **11**(1), 1–15. https://doi.org/10.1038/s41598-021-87598-w (2021).
30. Gao, S. J. *et al.* Prediction of distant metastases after stereotactic body radiation therapy for early stage NSCLC: Development and external validation of a multi-institutional model. *J. Thorac. Oncol.* **18**(3), 339–349. https://doi.org/10.1016/J.JTHO.2022.11.007 (2023).
31. Eriguchi, T. *et al.* Relationship between dose prescription methods and local control rate in stereotactic body radiotherapy for early stage non-small-cell lung cancer: Systematic review and meta-analysis. *Cancers* **14**, 15. https://doi.org/10.3390/CANCERS14153815/S1 (2022).
32. Senthi, S., Lagerwaard, F. J., Haasbeek, C. J. A., Slotman, B. J. & Senan, S. Patterns of disease recurrence after stereotactic ablative radiotherapy for early stage non-small-cell lung cancer: A retrospective analysis. *Lancet Oncol.* **13**(8), 802–809. https://doi.org/10.1016/S1470-2045(12)70242-5 (2012).
33. Velazquez, E. R. *et al.* Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci. Rep.* **3**, 1. https://doi.org/10.1038/SREP03529 (2013).
34. Van Griethuysen, J. J. M. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**(21), e104. https://doi.org/10.1158/0008-5472.CAN-17-0339 (2017).
35. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**(3), 297–302. https://doi.org/10.2307/1932409 (1945).
36. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**(2), 155. https://doi.org/10.1016/J.JCM.2016.02.012 (2016).
37. Nickerson, C. A. E. A note on 'a concordance correlation coefficient to evaluate reproducibility'. *Biometrics* **53**(4), 1503–1507 [Online]. Available: https://www.jstor.org/stable/2533516 (1997).
38. Kim, J. & Lee, J. H. A novel graphical evaluation of agreement. *BMC Med. Res. Methodol.* **22**(1), 1–9. https://doi.org/10.1186/S12874-022-01532-W/FIGURES/5 (2022).
39. Lin, L. I. K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**(1), 255–268. Available: https://www.jstor.org/stable/2532051 (1989).
40. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**(3–4), 591–611. https://doi.org/10.1093/BIOMET/52.3-4.591 (1965).
41. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**(2), 65–70 (1979).
42. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18**(1), 50–60. https://doi.org/10.1214/AOMS/1177730491 (1947).
43. Ding, C., & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *Bioinform. Comput. Biol.* **3**(2), 185–205. Accessed: Oct. 17, 2023. [Online]. Available: http://www.nersc.gov/~cding/MRMR/ (2005).
44. Tomek, I. Two modifications of CNN. *IEEE Trans. Syst. Man Cybern.* **6**(11), 769–772. https://doi.org/10.1109/TSMC.1976.4309452 (1976).
45. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
46. Chen, Y. & Yang, Y. The one standard error rule for model selection: Does it work?. *Stats (Basel)* **4**(4), 868–892. https://doi.org/10.3390/STATS4040051 (2021).
47. Akaike, H. Information theory and an extension of information the maximum theory likelihood and an principle extension of the maximum likelihood principle. *Biogeochemistry* **1998**, 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15/COVER (1998).
48. Reiazi, R. *et al.* The impact of the variation of imaging parameters on the robustness of Computed Tomography radiomic features: A review. *Comput. Biol. Med.* **133**, 104400. https://doi.org/10.1016/J.COMPBIOMED.2021.104400 (2021).
49. Varghese, B. A. *et al.* Identification of robust and reproducible CT-texture metrics using a customized 3D-printed texture phantom. *J. Appl. Clin. Med. Phys.* **22**(2), 98. https://doi.org/10.1002/ACM2.13162 (2021).
50. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 1. https://doi.org/10.1038/NCOMMS5006 (2014).
51. van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging—'how-to' guide and critical reflection. *Insights Imaging* **11**(1), 1–16. https://doi.org/10.1186/S13244-020-00887-2/TABLES/3 (2020).
52. Zhovannik, I. *et al.* Learning from scanners: Bias reduction and feature correction in radiomics. *Clin. Transl. Radiat. Oncol.* **19**, 33–38. https://doi.org/10.1016/J.CTRO.2019.07.003 (2019).

Author contributions

FL collected and curated the data. FL and FC annotated the data. TL established the radiomic workflow and the robustness assessment protocol. TL analyzed and interpreted the data. TL wrote the first draft of the manuscript. PL and RH supervised the project and revised the manuscript. All authors contributed to the manuscript and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-58551-4.

Correspondence and requests for materials should be addressed to T.L. or F.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.