Statistical harmonization can improve the development of a multicenter CT based radiomic model predictive of non-response to induction chemotherapy in laryngeal cancers

Ingrid Masson, Ronrick Da-Ano, François Lucia, Mélanie Doré, Joel Castelli, Camille Goislard de Monsabert, Jean-François Ramée, Selima Sellami, Dimitris Visvikis, Mathieu Hatt, et al.

Article type : Research Article

Article type: Research article

Title: Statistical harmonization can improve the development of a multicenter CT based radiomic model predictive of non-response to induction chemotherapy in laryngeal cancers

Short title: CT radiomic model in laryngeal cancer

Ingrid Masson [1], MD ; Ronrick Da-ano [1], MSc ; François Lucia [1,2,] MD ; Mélanie Doré [3], MD ; Joel Castelli [4,5], MD, PhD ; Camille Goislard de Monsabert [4], MD ; Jean-François Ramée [6], MD ; Selima Sellami [2,7], MD ; Dimitris Visvikis [1], PhD ; Mathieu Hatt [1], PhD ; Ulrike Schick [1,2], MD ; PhD

1 LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France

2 Radiation Oncology Department, University Hospital, Brest, France

3 Department of Radiation Oncology, Institut de cancérologie de l'Ouest René-Gauducheau, Saint-Herblain, France

4 Radiotherapy Department Cancer, Institute Eugène Marquis, Rennes, France

5 University of Rennes 1, LTSI, Rennes, France

6 Department of Medical Oncology, Centre Hospitalier de Vendée, La Roche sur Yon, France

7 Radiotherapy Department, Centre Hospitalier de Cornouaille, Quimper, France

Corresponding author:

Ingrid Masson

Laboratory of Medical Information Processing LaTIM INSERM UMR 1101, ACTION team (Therapeutic ACTion guided by multimodality Imaging in Oncology)

IBRBS (Institut Brestois de la Recherche en Biologie-Santé)

22 rue Camille Desmoulins, 29238 Brest, France

*Permanent address :
Department of Radiation Oncology
Institut de cancérologie de l'Ouest René-Gauducheau,
Boulevard Professeur Jacques Monod, 44805 Saint-Herblain, France
E mail address: ingrid.masson.im@gmail.com
Phone number: +33623475642

Abstract:

Purpose: To develop a radiomic model predicting non-response to induction chemotherapy in laryngeal cancers, from multicenter pre-therapeutic contrast-enhanced computed tomography (CE-CT) and evaluate the benefit of features harmonization in such a context.

Methods: Patients (n=104) eligible for laryngeal preservation chemotherapy were included in 5 centers. Primary tumor was manually delineated on the CE-CT images. The following radiomic features were extracted with an in-house software (MIRAS v1.1, LaTIM UMR 1101): intensity, shape and textural features derived from Grey Level Co-occurrence Matrix: GLCM; Neighbourhood Grey Tone Difference Matrix: NGTDM; Grey Level Run Length Matrix: GLRLM; Grey Level Size Zone Matrix: GLSZM. Harmonization was performed using ComBat after unsupervised hierarchical clustering, used to determine labels automatically, given the high heterogeneity of imaging characteristics across and within centers. Patients with similar features distributions were grouped with unsupervised clustering into an optimal number of clusters (2) determined with 'silhouette scoring'. Statistical harmonization was then carried out with ComBat on these two identified 2 clusters. The cohort was split into training/validation (n=66) and testing (n=32) sets. Area under the receiver operating characteristics curves (AUC) were used to evaluate the ability of radiomic features (before and after harmonization) to predict non-response to chemotherapy, and specificity (sp) and sensitivity (se) were used to quantify their performance in the testing set.

Results: Without harmonization, none of the features identified as predictive in the training set remained significant in the testing set. After ComBat, one textural feature identified in the training set keeps a predictive trend in the testing set: Zone Percentage, derived from the GLSZM, was predictive of non-response in the training set (AUC=0.62, Se=70%, Sp=64%, p=0.04) and obtained a satisfactory performance in the testing set (Se=80%, Sp=67%, p=0.03), although significance was limited by the size of the testing set. These results are consistent with previously published findings in head and neck cancers."

Conclusions: Radiomic features from CE-CT could help in the selection of patients for induction chemotherapy in laryngeal cancers, with relatively good sensitivity and specificity in predicting lack of response. Statistical harmonization with ComBat and unsupervised clustering seems to improve the predictive value of features extracted in such a heterogeneous multicenter setting.

Keywords: Imaging biomarkers and radiomics; CT; larynx cancer; prediction of treatment response; ComBat; unsupervised learning

Introduction

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide, accounting for 6% of all cancer cases [1]. Laryngeal cancers are among the most common and are mainly caused by tobacco and alcohol. The management of locally advanced laryngeal tumours in previously untreated patients is mainly based on laryngeal preservation strategies, two of which have been validated so far: induction chemotherapy followed by radiotherapy (RT) alone (GORTEC 2000-01 [2]) and RT with concurrent cisplatin (RTOG 91-11 [3]).

The rationale of the first option is to employ induction chemotherapy to select patients for subsequent treatment according to tumor response: either RT in responders or salvage laryngectomy followed by adjuvant RT in non responders. The first international phase III trials showed that larynx could be preserved in 40% – 60% of the patients with such an approach without compromising survival [4,5].

The standard induction chemotherapy regimen is a combination of docetaxel, cisplatin and 5-fluorouracil (TPF), which has demonstrated its superiority over cisplatin-5-fluorouracil (PF) in terms of overall response and larynx preservation rate in the French 2000-01 GORTEC (*Groupe Oncologie Radiothérapie Tête Et Cou*) trial [2]. Nevertheless, the toxicity of the TPF regimen is increased compared to a FP regimen, with 31.5% *vs.* 17.6% of grade 4 neutropenia, respectively. According to the results of the GORTEC 2000-01 study, 20% of patients do not respond to induction chemotherapy with TPF, and are therefore unnecessarily exposed to significant toxicity when they will eventually have to undergo laryngectomy anyway. There is currently no established pre-treatment biomarker to identify these patients.

These patients usually undergo contrast-enhanced computed tomography (CE-CT) as a clinical routine diagnostic imaging procedure. Our first goal was to investigate if radiomic features from CE-CT could contribute in predicting lack of response to TPF chemotherapy. In order to build a more relevant and generalizable model, we aimed for a multicenter recruitment. In that context, it is known that CT radiomic features exhibit variable levels of sensitivity on scanner model/manufacturer, as well as acquisition and reconstruction settings (including slice thickness, matrix size, tube current, etc.) [6–9]. This is particularly problematic when CT images come from different centers relying on different machines and protocols. Our second goal was therefore to evaluate the benefit of features harmonization in that context.

Methods

Patients

All patients with histologically proven locally advanced laryngeal or hypopharyngeal cancer, who were treated with laryngeal preservation using TPF induction chemotherapy from June 2008 to January 2018 at 5 French institutions (Brest, Nantes, Rennes, La Roche-sur-Yon and Quimper), were retrospectively considered. Only patients with a Performance Status (PS) 0-1 were included. The disease had to be curable with total (pharyngo)laryngectomy and postoperative radiotherapy. A direct laryngoscopy was required to assess laryngeal mobility. Patients were considered for laryngeal preservation in case of T2 laryngeal tumor not accessible to partial laryngectomy, or T3 tumor without massive infiltration of the endolarynx. Nodal status ranged from N0 to N3. Patients with transglottic T3 tumours with massive hemilaryngeal infiltration or T4 with massive cartilage invasion or tumours of the retrocricoarythenoidal region or posterior hypopharyngeal wall were not included. Patients with a tumour requiring initial tracheotomy, or accessible for partial surgery or requiring circular hypopharyngeal surgery, were also excluded.

This study was approved by the local ethics committee (29BRC19.0006) and all patients gave their consent via a non-opposition form.

Treatment and main outcome

Two or three cycles of TPF chemotherapy were planned according to each center's protocol. Response to chemotherapy was assessed after 2 or 3 cycles depending on centers and was based on clinical examination (endoscopic evaluation of larynx mobility) and imaging evaluation (computed tomography (CT) scan and/or [18F]-Fluoro-Deoxy-Glucose (18F-FDG) Positron Emission Tomography (PET)/CT). Primary endpoint was lack of response to chemotherapy, defined as a non-remobilization of the larynx if laryngeal mobility was decreased or abolished at diagnosis, or a response of the primary tumor <50% (RECIST criteria). Non-responders were referred for a (pharyngo-)laryngectomy followed by postoperative RT, whereas responders received conservative RT with or without chemotherapy depending on pathological risk factors.

Image acquisition and definition of volume of interests

All patients had a CE-CT at diagnosis. A great variability existed in terms of scanner models, collimation / acquisition settings and reconstruction parameters, even within a single centre (details provided in Supplemental Table 1). For each patient, the volume of interest (VOI) of the primary tumor was manually delineated on the axial slices of the pre-therapeutic CE-CT by one experimented radiation oncologist (IM), using the open source software 3D slicer v.4.11 (http://www.slicer.org/).

Image interpolation

There was a wide variety of voxel sizes and slice thickness in the original images (see Supplemental Table 1). We chose not to interpolate images to a common voxel size to avoid investigating several different interpolation methods and different target dimensions, as each approach could lead to artifacts and impact the interpolated images in different ways. All images were therefore processed in their native dimensions.

Intensity discretization and features extraction

IBSI (image biomarker standardisation initiative) compliant [10] radiomic features (intensity, shape and textural) were extracted with an in-house software (MIRAS v1.1, LaTIM UMR 1101). Textural features were derived from 4 matrices representing the spatial distribution of voxel intensities at local (GLCM: Grey Level Co-occurrence Matrix; NGTDM: Neighbourhood Grey Tone Difference Matrix) and regional (GLRLM: Grey Level Run Length Matrix; GLSZM: Grey Level Size Zone Matrix) scales. Texture matrices were implemented in 3D following the merging strategy (i.e., considering all 13 directions simultaneously). Four different discretization of voxels intensity values were implemented prior to textural features extraction through three methods[11]: fixed bin size (FBS) with either 10 or 25 Hounsfield Units (HU) [12,13], fixed bin number (FBN) with 64 bins [14] and histogram equalization (HE) into 64 bins [15,16]. These have been described as providing a good compromise between information on image heterogeneity and noise. As the various discretization techniques have different advantages and drawbacks, we chose to use them all in order to benefit from the texture optimization process (i.e., certain features might provide more relevant information using a specific discretization scheme, thus if using only one discretization approach, part of the features could not be as informative as they could be) [16,17]. Note that HE is not part of the IBSI standardization yet, so textural features obtained with this discretization cannot be considered IBSI-compliant.

A total of 274 radiomic variables were thus evaluated: 15 shape features, 11 first-order statistical features and 62 textural features calculated with the 4 discretization settings mentioned above (Supplemental Table 2).

Harmonization method for multicenter data

To correct for the high variability of the acquisition settings and reconstruction parameters (Supplemental Table 1) and their well-known impact on most radiomic features distributions, we used the statistical harmonization method ComBat, initially developed for genomics to correct for batch effects [18]. It removes inter-site technical variability while preserving biological variability. It has been successfully applied to multicenter radiomics on CT [19], magnetic resonance imaging (MRI) [20] and PET [20,21] images. We used the ComBat parametric model with its R implementation, available at the following address: https://github.com/Jfortin1/ComBatHarmonization/tree/master/R. Harmonization was performed on all previously extracted radiomic features. We did not include a biological covariate in our computation given the high homogeneity of our population in terms of clinical or histological data. ComBat requires grouping patients whose scans have been performed with the same settings and although ComBat was shown to be robust for small samples, it is nonetheless recommended that these groups should contain enough

patients for the estimation to be performed. Because of the very high heterogeneity in our cohort, using ComBat directly would have led to use more than 15 labels, with labels having as few as one or two patients. Therefore, we rather chose to first perform unsupervised hierarchical clustering to group patients with sufficiently similar features distributions. Hierarchical clustering is a type of unsupervised algorithm which groups data by similarity – it classifies objects without any prior knowledge of the class they belong to, based on the measure of the Euclidean distance (**Figure 1a.**) [22]. To determine the optimal number of clusters to consider before running the hierarchical clustering, we used 'silhouette' scoring, a tool used to validate the clustering [23]. The 'silhouette' is then constructed to determine the optimal number of cluster with a ratio scale data (as in the case of Euclidean distance) that is suitable for clearly separated clusters (Supplemental Figure 1) [23]. The assumption made here is that the imaging differences have a stronger impact on the distribution of radiomic features compared to the ones we need to capture in order to classify patients with respect to lack of response to therapy. As there was an obvious risk that the unsupervised clustering could end up grouping patients based on clinical endpoint rather than imaging differences, the resulting clusters were checked for consistency regarding their percentage of non-responders (**Figure 1b.**). To increase the confidence in the resulting clusters, the exact same technique was also applied to another multicenter dataset (not exploited further here) of 197 cervical cancer patients from 3 centers [20,24], where the true labels are known.

Statistical analysis

Stratified sampling was used: patients were sorted in chronological order according to the date of diagnosis and then split into a training/validation set (2/3) and a testing (1/3) set. The comparison of the training/validation and testing sets was tested under the null hypothesis H0 with a risk of $\alpha = 0.05$. The discrete quantitative variables (age and BMI) were tested with the Mann-Whitney U test. The qualitative variables were tested with the Chi2 parametric test.

The evaluation and selection of relevant variables, as well as the identification of a threshold value for optimizing sensitivity and specificity in identifying lack of response was carried out in the training/validation set. No correction for multiple testing was carried out for the discovery of potentially predictive features in the training/validation set, as the final evaluation of their statistical significance lies in their performance in the testing set. The 274 radiomic features values, 8 clinical variables (sex, age, PS, BMI, tumor location, T stage, lymph nodes involvement, pre-therapeutic mobility of the larynx) and 1 treatment parameter (number of TPF cycles) were tested for their ability to predict lack of response to induction chemotherapy, as defined above. The predictive performance was quantified in univariate

analysis by the AUC of the ROC curves. Optimal cut-off values for each variable were defined using the Youden Index. Selection of radiomic features to be evaluated in the testing set was then performed based on 3 criteria: an AUC ≥ 0.60, a minimum specificity of 60% (arbitrary chosen to avoid false positive events) and a lack of redundancy between the selected parameters, based on Spearman's rank correlation coefficients. Spearman's rank correlation between two features was considered significant (with 0 = no correlation; -1 = negative correlation; +1 = positive correlation) at a significance level alpha = 0.05. In case of such redundancy, the most predictive feature in the training set was kept for evaluation in the testing set. The threshold of 60% specificity was chosen based on a clinical rationale, in order to limit the number of patients wrongly classified as non-responders (and for whom the sentence is total laryngectomy), when in fact they are responders

The previously selected variables and/or combinations of variables (with their optimal threshold) were then evaluated in the testing set. The Bonferroni method was used to correct for significance for multiple testing comparisons in the testing set. Significance was defined as a corrected p value below α/n with $\alpha=0.05$ and n=the number of tests performed

The statistical analyses were performed using MedCalc Statistical Software version 15.8 (MedCalc Software bvba, Ostend, Belgium; https://www.medcalc.org; 2015).

Results

Patient characteristics

One hundred and four patients were included between June 2008 and January 2018. Ninety-eight patients were analyzed (flow chart in Supplemental Figure 2). Patients in the training/validation (n=66) and testing (n=32) sets had similar clinical and pathological characteristics (**Table 1**). Overall, 10 patients (15%) and 5 patients (16%) did not respond to TPF induction chemotherapy in the training/validation and testing sets, respectively.

Without ComBat harmonization

In the training/validation set and in univariate analysis, one clinical variable (gender) and 33 radiomic features were significantly correlated with lack of response to chemotherapy (without correction for multiple testing). After selection of features according to the criteria defined above, no clinical variable (Supplemental Table 3) and only 1 radiomic features was predictive enough and non-redundant to be retained for further evaluation in the testing set : Large Area Low Gray Level Emphasis with discretization

using FBS with 25 HU (LALGLE $_{GLSZM, FBS:25HU}$) (**Figure 2**) reached AUC 0.680 (p<0.001). The optimal threshold for LALGLE $_{GLSZM, FBS:25HU}$ in the training set was ≤12, resulting in 80% sensitivity and 70% specificity.

LALGLE $_{GLSZM, FBS:25HU}$ with its thresholds demonstrated however low predictive power in identifying non-responders in the testing set with Se and Sp of 80% and 52% (p=0.15), respectively.

With ComBat harmonization

When applied to the 197 cervical cancer patients cohorts with 3 known labels, the unsupervised clustering almost perfectly (only 1 patient misclassified) labelled all patients from the 3 different centers [20,24]. In the present cohort, the unsupervised clustering automatically identified two clusters of 38 and 60 patients (**Figure 1a.**), which exhibited similar proportion of events: 6 non-responders for 38 patients (16%) in the first cluster and 9 non-responders for 60 patients (15%) in the other (**Figure 1b.**). It is thus very unlikely that this unsupervised differentiation was based on outcome, rather than on differences of features due to imaging. These two clusters were therefore used as labels for the ComBat harmonization.

In the training/validation set and in univariate analysis, 27 radiomic features were significantly correlated with lack of response to chemotherapy after this harmonization based on labels determined through unsupervised clustering. After selection of features according to the criteria defined above, 3 were kept, all derived from GLSZM with discretization using FBS with 25 HU : Zone Percentage (ZP $_{GLSZM, FBS:25HU}$) with AUC = 0.62 and a threshold of 0.07 (Se = 70%, Sp = 64%, p=0.04), Large Area Low Gray Level Emphasis (LALGLE $_{GLSZM, FBS:25HU}$) with AUC = 0.687 and a threshold of ≤-103 (Se = 60%, Sp = 82%, p = 0.01) and Large Area High Gray Level Emphasis (LAHGLE $_{GLSZM, FBS:25HU}$) with AUC = 0.641 and a threshold of ≤9283393 (Se = 70%, Sp = 68%, p=0.02). The 3 textural features were not correlated with each other (**Figure 3a**).

In the testing set, Zone Percentage with its threshold was the only feature able to predict non-responders with Se = 80% and Sp = 67%, CI 95% [55%-88%], p=0.03 (**Figure 3b**) although this was not significant anymore after correction for multiple testing (threshold at p=0.017). LALGLE and LAHGLE with their respective thresholds failed to predict non responders in the testing set with Se and Sp of 80%, 41% (p=0.4) and 60%, 63% (p=0.4), respectively.

Table 2 is a sample of the differences between the radiomic features values before and after harmonization by ComBat in the training set, as well as the optimal cut-off according to the Youden index. The textural features with the highest AUCs for the prediction of non-response to chemotherapy are compared through DeLong et al., method (1988) in **Figure 4**. As explained above, 27 radiomic features after harmonization were significantly correlated with lack of response to TPF in the training set and in

univariate analysis, such as Short run high grey level emphasis (SRHGLE $_{GLRLM, FBS:25HU}$) and Large Area High Gray Level Emphasis (LAHGLE $_{GLSZM, FBS:25HU}$); but they have not been retained for evaluation in the testing set due to insufficient specificity/AUC or to a correlation with other features. The combination of the most promising features with the Zone Percentage in the training set also failed to create a superior radiomic model, in terms of AUC and specificity, compared to the Zone Percentage alone. No combination of radiomic features was therefore evaluated in the testing set.

Performance of radiomics-based models was compared against tumor volume, as recommended by Vallieres, *et al* [17,25]. ROC curve for tumor volume did not show any predictive ability of non-response to TPF in the training set (AUC 0.52, CI 95% [39%-64%]), therefore tumor volume could not be retained for evaluation in the testing set. Tumor volume and Zone Percentage showed a negative moderate correlation with a Spearman rank coefficient = -0.23 when considering the entire set of 98 patients (p=0.021, CI 95% [-0.41-0.036]).

Discussion

The rate of non-response to induction TPF (15-16%) reported here is very similar to the one of the GORTEC 2000-01 trial with 20% of non-responders in the TPF group [2]. None of the clinical variables (age, gender, PS, BMI, site of tumor, T stage, lymph nodes status, pretherapeutic hemilaryngeal mobility) or treatment modalities (2 *vs* 3 TPF cyles) were found to be statistically significant in identifying non-responders to chemotherapy. This might be due to the high homogeneity of eligible patients for laryngeal preservation protocol. In the literature, only few variables have been reported as predictors of lack of response to organ preservation strategy, namely a nonfunctional larynx (extensive T3 or T4a) or tumor invasion through cartilage into surrounding soft tissues [2,26,27]. The fact that tumor volume and stage did not emerge as predictors of lack of response to TPF in our cohort could be related to the small number of patients in each subgroup (2 and 1 patients with T4a stage in the training and testing sets, respectively). Patients with cartilage invasion or large volume disease are poor candidates for laryngeal preservation [26]. They may respond to conservative treatment, which is consistent with our study, but have a high local early recurrence rate [28]. The number of TPF cures (2 vs. 3) also did not emerge as a predictor of lack of response. Given the fact that no clinical variable was associated with the endpoint studied, we could not compare our radiomic findings with a basic clinical model, nor build a combined radiomic - clinical one, as previously done by others [17].

The analysis performed on the raw features (i.e., without multicenter harmonization) did not allow for features selected in the training/validation set to be successfully evaluated in the testing set. This was expected as the features are derived from images exhibiting very high variability in terms of scanner models, acquisition protocols and reconstruction parameters. Indeed, in this multicenter study, which was extended over 10 years, 26 different machines were used to produce the CE-CT. In addition, for the same machine, the acquisition and reconstruction protocols varied. Berenguer, *et al*. showed that most CT radiomic features are not reproducible when changing the pitch factor or the reconstruction kernel (intra-CT analysis) [7]. Reproducibility of inter-CT radiomic features was also poor when comparing different scanners with the same settings. Shafiq-ul-Hassan, *et al*. studied the reproducibility of 213 radiomic features by changing the slice thickness and pixel size (FOV): 42 out of 213 parameters showed significantly better reproducibility after normalizing the voxel size [6]. Reconstructed section thickness and reconstruction kernel can also infer high variability [29].

In contrast, the selection of a radiomic feature amongst these harmonized with ComBat allowed identifying one that met our selection criteria based on AUC, specificity and absence of redundancy, and led to some predictive ability (with a trend) in the testing set : Zone Percentage. Two other features (LALGLE and LAHGLE) exhibited good performance in the training set, but failed to predict non responders in the testing set. The performance of the radiomic model was not improved by combining Zone Percentage with these 2 other features. The combinations were tested manually. This prevented us from exploring all possible combinations, compared to a more exhaustive machine learning approach. Zone percentage is a textural feature derived from the GLSZM that characterizes texture at a regional level (groups of voxels). Highly uniform regions of interest produce a low zone percentage [10]. These results are consistent with the literature. Bogowicz *et al*. already demonstrated that greater heterogeneity in pre-treatment CT images of patients with head and neck cancer is associated with poor local tumour control after definitive chemoradiotherapy [30]. The underlying hypothesis is that this heterogeneity on imaging reflects the heterogeneity of the microenvironment (vascularization, necrosis, tumor hypoxia), which is known to be associated with a more aggressive tumor phenotype in head neck cancers [31]. For the same reason, CE-CT might be more informative for radiomics than non-injected CT [32].

Four different discretization settings of voxel intensities were implemented for each textural feature. The selected feature Zone Percentage was predictive when calculated using the fixed bin size method with 25 HU. This discretization approach has been shown to produce more robust radiomic features compared to a fixed number of bins [12,33,34]. In addition, most of the textural features identified above as potential

predictive factors, including Zone Percentage, were derived from a discretization of intensity values over equally spaced 25-HU bins, rather than 10 (Table 2).

In our study, a number of radiomic features (33 before and 27 after harmonization) were correlated with lack of response to chemotherapy. It was necessary to carry out a step of features selection since they have a variable level of rank intercorrelation. Several strategies exist [15,19,35,36]. We have deliberately chosen a simple one based on the search for the highest possible specificity to be clinically relevant: a high specificity and therefore a low rate of false positives would deprive as few responders as possible of an attempt to preserve their larynx.

There are several limitations in our study. First, our cohort is of limited size and retrospective. This is due to the restrictive eligibility criteria for the laryngeal preservation strategy [2] in clinical practice. Ideally, a larger population, with more events, would improve model training. Second, manual segmentation by a single radiation oncologist (IM) was performed, which is time-consuming and is a source of inter-operator variability [37]. However, manual segmentation is mainly used in head and neck cancers [32,35,38–40] as these tumors are less suitable for automatic or semi-automatic segmentation than other tumor locations. Moreover, the delineation of the GTV often requires the expertise of a radiation oncologist, as mainly based on clinical evaluation. Finally, Geets, *et al*. did not find significant inter-observer differences regarding the delineation of laryngeal tumours on the CT if consistent delineation guidelines are followed [41]. Pavic and al. also found an acceptable stability for inter-observer HNSCC delineation [42]. Another limitation concerns image interpolation. We chose not to interpolate images because it would mean evaluating several different methods and voxel target size, which would be extremely time-consuming. Interpolating images to a common voxel size may help in reducing the differences between the resulting radiomic features (as most of them have been shown to be dependent on voxel size), but has been shown to be insufficient, compared to an approach such as ComBat [43]. A full comparison of various interpolation techniques versus ComBat-type harmonization could be of interest but is out of the scope of the present work. Indeed, one challenge in our study was the lack of standardization in the acquisition and reconstruction parameters of CT data, corresponding to the clinical routine practice of the five centers where patients were retrospectively recruited. We used ComBat, which is a data harmonization method where “batch” effects are estimated *a posteriori* and removed from the data [18]. ComBat works well even for small samples, as long as the number of patients in each batch is about the same order of size, which is the case in our study. One of the risk of ComBat is that it may confuse biological and technical heterogeneity, thus reducing the test's ability to identify differences between responders and non-responders, as noted by Goh *et al* [44].

Another potential risk is related to the use of unsupervised hierarchical clustering to determine the labels. It was especially important to check that the resulting clusters were indeed defined based on differences due to imaging variability and not clinical outcome. Because the two obtained clusters had almost the same percentage of non-responders, it can be safely assumed that they were indeed the results of measured differences due to imaging acquisition and associated processing protocols, rather than different outcome profiles. In addition, when performing the same technique to another dataset with known labels, it misclassified only one patient out of 197, strengthening further our confidence in the resulting clusters [20]. However, even by grouping patients through unsupervised hierarchical clustering, it was not possible to group patients with a perfect match for all acquisition and reconstruction parameters. These limitations may partly explain the inability in our study to obtain textural features with higher AUC and specificities. Nevertheless, the variability of the CT protocols in our study is representative of current clinical practice and constitutes a major challenge for the possible implementation of radiomics in therapeutic decision-making.

Conclusion

Statistical harmonization can help for the development of a multicenter CT based radiomic model predictive of non-response to induction chemotherapy in laryngeal cancers. Without harmonization, performance for all investigated features failed to reach statistical significance, in a highly heterogeneous and multicenter setting. With harmonization one promising feature was identified, although its significance level was limited by the size of the testing set. These findings now require evaluation in an external cohort, which could then lead to larger, prospective validation.

Figure legends

Figure 1. **Unsupervised clustering approach**

1.a. Identifying number of patients based on separated clusters

1.b. Distribution of the responders and non-responders in terms of the clusters

1.c. Distribution of the responders and non-responders in the training and the testing sets respectively

Figure 2. Performance in the task of predicting non response to therapy, in the training set, of the most promising textural feature before harmonization.

ROC curves of Large Area Low Gray Level Emphasis (LALGLE $_{GLSZM, FBS:25HU}$) The Youden index is represented on the ROC curve by a white dot. The values corresponding to the Youden index (associated criterion, sensitivity (Se) and specificity (Sp), significance level p) are also given.

Figure 3.

3a. Correlation matrix (Spearman) between ZP $_{GLSZM, FBS:25HU}$; LALGLE $_{GLSZM, FBS:25HU}$; LAHGLE $_{GLSZM, FBS:25HU}$

Values in bold were significantly different from 0 at a significance level alpha = 0.05

3b. Performance in the task of predicting non response to therapy, of Zone Percentage after harmonization.

ROC curve of Zone Percentage (ZP $_{GLSZM, FBS:25HU}$). This textural feature was identified as the best biomarker for predicting non-responders in the training set and led to some predictive ability in the testing set, although it was only a trend failing to reach statistical significance after correction for multiple testing.

Figure 4. Comparison of ROC curves in the training set of the most promising radiomic features in terms of AUC, for the prediction of non response to therapy. DeLong et al., method (1988) was applied. Results are presented before (Fig 4a.) and after (Fig 4b.) ComBat harmonization.

Data Sharing and Data Accessibility

Data available on request due to privacy/ethical restrictions. The data that support the findings of this study are available on request from the corresponding author, [IM]. The data are not publicly available due to privacy and ethical restrictions.

References

1. John Andrew Ridge MD. Head and Neck Tumors. Cancer Network. Published June 2, 2016. Accessed August 5, 2019. https://www.cancernetwork.com/cancer-management/head-and-neck-tumors

2. Pointreau Y, Garaud P, Chapet S, et al. Randomized Trial of Induction Chemotherapy With Cisplatin and 5-Fluorouracil With or Without Docetaxel for Larynx Preservation. *J Natl Cancer Inst*. 2009;101(7):498-506. doi:10.1093/jnci/djp007
3. Forastiere AA, Goepfert H, Maor M, et al. Concurrent Chemotherapy and Radiotherapy for Organ Preservation in Advanced Laryngeal Cancer. *New England Journal of Medicine*. 2003;349(22):2091-2098. doi:10.1056/NEJMoa031317
4. Department of Veterans Affairs Laryngeal Cancer Study Group, Wolf GT, Fisher SG, et al. Induction chemotherapy plus radiation compared with surgery plus radiation in patients with advanced laryngeal cancer. *N Engl J Med*. 1991;324(24):1685-1690. doi:10.1056/NEJM199106133242402
5. Lefebvre JL, Chevalier D, Luboinski B, Kirkpatrick A, Collette L, Sahmoud T. Larynx preservation in pyriform sinus cancer: preliminary results of a European Organization for Research and Treatment of Cancer phase III trial. EORTC Head and Neck Cancer Cooperative Group. *J Natl Cancer Inst*. 1996;88(13):890-899. doi:10.1093/jnci/88.13.890
6. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44(3):1050-1062. doi:10.1002/mp.12123
7. Berenguer R, Pastor-Juan M del R, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*. 2018;288(2):407-415. doi:10.1148/radiol.2018172361
8. Espinasse M, Pitre-Champagnat S, Charmettant B, et al. CT Texture Analysis Challenges: Influence of Acquisition and Reconstruction Parameters: A Comprehensive Review. *Diagnostics*. 2020;10(5):258. doi:10.3390/diagnostics10050258
9. Kim H, Park CM, Lee M, et al. Impact of Reconstruction Algorithms on CT Radiomic Features of Pulmonary Tumors: Analysis of Intra- and Inter-Reader Variability and Inter-Reconstruction Algorithm Variability. *PLOS ONE*. 2016;11(10):e0164924. doi:10.1371/journal.pone.0164924
10. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv:161207003 [cs]*. Published online December 21, 2016. Accessed April 1, 2019. http://arxiv.org/abs/1612.07003
11. Desseroit MC, Tixier F, Cheze Le Rest C. Comparison of three quantization methods for the calculation of textural features in PET/CT images: impact on prognostic models in non-small cell lung cancer. In: *IEEE NSS-MIC*. ; 2016.
12. Leijenaar RTH, Nalbantov G, Carvalho S, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Scientific Reports*. 2015;5:11075. doi:10.1038/srep11075

13. Leijenaar RT, Bogowicz M, Jochems A, et al. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. *BJR*. 2018;91(1086):20170498. doi:10.1259/bjr.20170498
14. Bogowicz M, Riesterer O, Bundschuh RA, et al. Stability of radiomic features in CT perfusion maps. *Phys Med Biol*. 2016;61(24):8736-8749. doi:10.1088/1361-6560/61/24/8736
15. Lucia F, Visvikis D, Desseroit M-C, et al. Prediction of outcome using pretreatment18F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy. *Eur J Nucl Med Mol Imaging*. Published online December 9, 2017. doi:10.1007/s00259-017-3898-7
16. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol*. 2015;60(14):5471-5496. doi:10.1088/0031-9155/60/14/5471
17. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep*. 2017;7. doi:10.1038/s41598-017-10371-5
18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037
19. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology*. 2019;291(1):53-59. doi:10.1148/radiol.2019182023
20. Lucia F, Visvikis D, Vallières M, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *Eur J Nucl Med Mol Imaging*. Published online December 7, 2018. doi:10.1007/s00259-018-4231-9
21. Orlhac F, Boughdad S, Philippe C, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *J Nucl Med*. 2018;59(8):1321-1328. doi:10.2967/jnumed.117.199935
22. Murtagh F, Contreras P. Methods of Hierarchical Clustering. *arXiv:11050121 [cs, math, stat]*. Published online April 30, 2011. Accessed September 5, 2019. http://arxiv.org/abs/1105.0121
23. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7
24. Da-ano R, Masson I, Lucia F, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Scientific Reports*. 2020;10(1):10248. doi:10.1038/s41598-020-66110-w
25. Vallieres M, Zwanenburg A, Badic B, Cheze-Le Rest C, Visvikis D, Hatt M. Responsible Radiomics Research for Faster Clinical Translation. *J Nucl Med*. Published online November 24, 2017. doi:10.2967/jnumed.117.200501

26. Forastiere AA, Weber RS, Trotti A. Organ Preservation for Advanced Larynx Cancer: Issues and Outcomes. *J Clin Oncol*. 2015;33(29):3262-3268. doi:10.1200/JCO.2015.61.2978
27. Forastiere AA, Ismaila N, Lewin JS, et al. Use of Larynx-Preservation Strategies in the Treatment of Laryngeal Cancer: American Society of Clinical Oncology Clinical Practice Guideline Update. *J Clin Oncol*. 2018;36(11):1143-1169. doi:10.1200/JCO.2017.75.7385
28. Patel UA, Howell LK. Local response to chemoradiation in T4 larynx cancer with cartilage invasion. *The Laryngoscope*. 2011;121(1):106-110. doi:10.1002/lary.21181
29. Meyer M, Ronald J, Vernuccio F, et al. Reproducibility of CT Radiomic Features within the Same Patient: Influence of Radiation Dose and CT Reconstruction Settings. *Radiology*. 2019;293(3):583-591. doi:10.1148/radiol.2019190928
30. Bogowicz M, Riesterer O, Ikenberg K, et al. Computed Tomography Radiomics Predicts HPV Status and Local Tumor Control After Definitive Radiochemotherapy in Head and Neck Squamous Cell Carcinoma. *International Journal of Radiation Oncology*Biology*Physics*. 2017;99(4):921-928. doi:10.1016/j.ijrobp.2017.06.002
31. Caudell JJ, Torres-Roca JF, Gillies RJ, et al. The future of personalised radiotherapy for head and neck cancer. *The Lancet Oncology*. 2017;18(5):e266-e273. doi:10.1016/S1470-2045(17)30252-8
32. Bogowicz M, Tanadini-Lang S, Veit-Haibach P, et al. Perfusion CT radiomics as potential prognostic biomarker in head and neck squamous cell carcinoma. *Acta Oncologica*. 2019;0(0):1-5. doi:10.1080/0284186X.2019.1629013
33. Yip SS, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol*. 2016;61(13):R150-R166. doi:10.1088/0031-9155/61/13/R150
34. Lovinfosse P, Visvikis D, Hustinx R, Hatt M. FDG PET radiomics: a review of the methodological aspects. *Clin Transl Imaging*. 2018;6(5):379-391. doi:10.1007/s40336-018-0292-9
35. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*. 2014;5:4006. doi:10.1038/ncomms5006
36. Larue RTHM, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *BJR*. 2016;90(1070):20160665. doi:10.1259/bjr.20160665
37. Hermans R, Feron M, Bellon E, Dupont P, Bogaert WV den, Baert AL. Laryngeal tumor volume measurements determined with CT: A study on intra- and interobserver variability. *International Journal of Radiation Oncology • Biology • Physics*. 1998;40(3):553-557. doi:10.1016/S0360-3016(97)00853-5

38. Leijenaar RTH, Carvalho S, Hoebers FJP, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncologica*. 2015;54(9):1423-1429. doi:10.3109/0284186X.2015.1061214
39. Bagher-Ebadian H, Siddiqui F, Liu C, Movsas B, Chetty IJ. On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. *Medical Physics*. 2017;44(5):1755-1770. doi:10.1002/mp.12188
40. Bogowicz M, Riesterer O, Stark LS, et al. Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. *Acta Oncologica*. 2017;56(11):1531-1536. doi:10.1080/0284186X.2017.1346382
41. Geets X, Daisne J-F, Arcangeli S, et al. Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: Comparison between CT-scan and MRI. *Radiotherapy and Oncology*. 2005;77(1):25-31. doi:10.1016/j.radonc.2005.04.010
42. Pavic M, Bogowicz M, Würms X, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol*. 2018;57(8):1070-1074. doi:10.1080/0284186X.2018.1445283
43. Reuzé S, Orlhac F, Chargari C, et al. Prediction of cervical cancer recurrence using textural features extracted from 18 F-FDG PET images acquired with different scanners. *Oncotarget*. 2017;8(26):43169-43179. doi:10.18632/oncotarget.17856
44. Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*. 2017;35(6):498-507. doi:10.1016/j.tibtech.2017.02.012

Acknowledgments:

This work was supported by the Foundation for Medical Research (FRM); grant number: M2R201806006004

We thank Pr Malard and Dr Ferron, from Department of Otololaryngology (University Hospital of Nantes, Nantes, France) for their support for our work.

Conflicts of interest: None declared.

Table 1. **Patient characteristics**

	Training/validation set (n= 66)		Testing set (n=32)		
	n	%	n	%	**P value**
Median age in years (interquartile range, range)	59.5 (9, 33)		60.5 (10, 29)		0.417
Sex					0.792
H	57	86	27	84	
F	9	14	5	16	
PS					0.425
0	44	67	18	56	
1	21	32	14	44	
Missing data	1	1	0	0	
BMI					0.236
< 18.5	5	8	3	9	
18.5-24.9	32	48	10	31	
25-29.9	15	23	14	44	
30-39.9	10	15	4	13	
Missing data	4	6	1	3	
Site of tumor					0.213
Supra glottic larynx	19	29	13	41	
Glottic larynx	14	21	9	28	
Hypopharynx	33	50	10	31	
Primary tumor stage					0.576
T2	8	12	2	6	
T3 without cord fixation	22	33	8	25	
T3 with fixed cord involvment	34	52	21	66	
T4a	2	3	1	3	
Larynx mobility					0.684
Normal	24	36	9	28	
Reduced	13	20	8	25	
Abolished	29	44	15	47	
Nodal stage					0.158
N0	27	41	13	41	
N1	11	17	3	9	
N2a	7	11	3	9	
N2b	12	18	3	9	
N2c	8	12	5	16	

N3	1	2	4	13	
Missing data	0	0	1	3	
Number of TPF chemotherapy cycles					0.692
1	1	2	1	3	
2	7	10	2	6	
3	58	88	29	91	

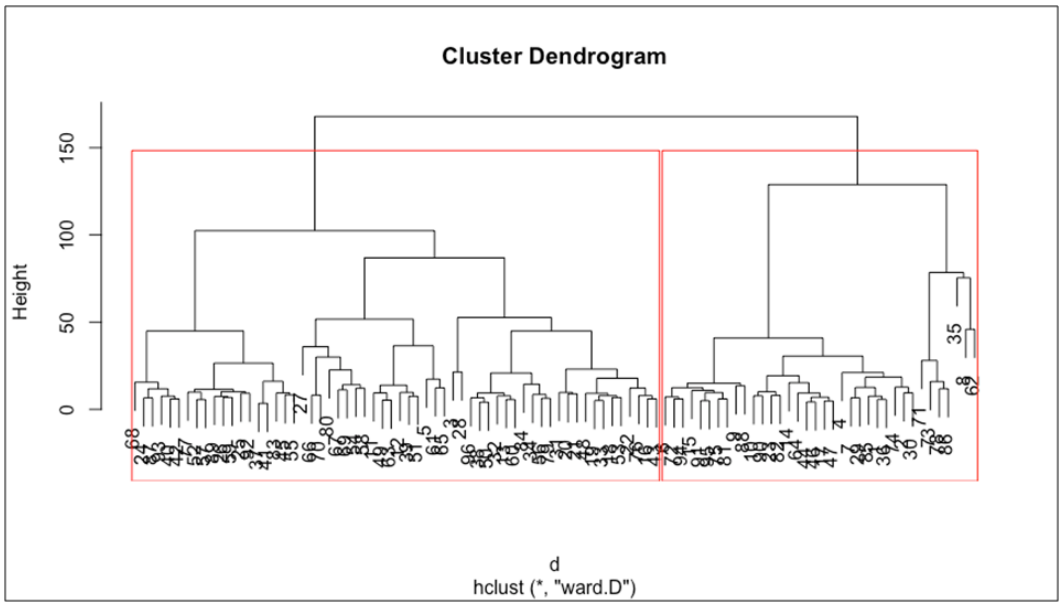
PS : Performans Status, BMI : Body Mass Index, TPF : Docetaxel Cisplatin Fluorouracil

Table 2. **Comparison of textural features before and after ComBat harmonization and their respective performance in predicting non response to therapy:** Some textural features with their optimal cut-off in the training/validation set, as well as their predictive performance evaluated in the testing set, are presented. More features successfully predict non-response to chemotherapy in the testing set after harmonization. Zone percentage (in bold) was the only feature that showed a trend towards some predictive ability in the testing set.

	Before harmonization								After harmonization							
	Training/validation set					Testing set			Training/validation set					Testing set		
	cut-off	AUC and 95% CI	Se	Sp	p	Se	Sp	p**	cut-off	AUC and 95% CI	Se	Sp	p	Se	Sp	p**
sum average $_{GLCM, FBS:10HU}$	>78	0.596 [47%-72%]	90	39	0.0145	*	*	*	>115	0.584 [46%-70%]	80	50	0.0447	*	*	*
autocorrelation $_{GLCM, FBS:25HU}$	>256	0.596 [47%-72%]	90	39	0.0145	*	*	*	>609	0.582 [45%-70%]	80	50	0.0447	*	*	*
joint average $_{GLCM, FBS:25HU}$	>15	0.596 [47%-72%]	90	39	0.0145	*	*	*	>23	0.584 [46%-70%]	80	50	0.0447	*	*	*
sum average $_{GLCM, FBS:25HU}$	>31	0.596 [47%-72%]	90	39	0.0145	*	*	*	>46	0.584 [46%-70%]	80	50	0.0447	*	*	*
cluster shade $_{GLCM, FBN:64}$	>-197	0.609 [48%-73%]	80	61	0.0062	*	*	*	>253	0.546 [42%-67%]	40	87	0.1043	*	*	*
SRHGLE $_{GLRLM, FBS:25HU}$	>147	0.596 [47%-72%]	90	41	0.0096	*	*	*	>366	0.577 [45%-70%]	80	55	0.0178	*	*	*
HGLRE $_{GLRLM, FBS:25HU}$	>249	0.595 [47%-72%]	90	40	0.0145	*	*	*	>603	0.580 [45%-70%]	80	50	0.0447	*	*	*
SAHGLE $_{GLSZM, FBS:25HU}$	>106	0.607 [48%-73%]	90	43	0.0063	*	*	*	>256	0.596 [47%-72%]	80	57	0.0127	*	*	*
ZP $_{GLSZM, FBS:25HU}$	>0.07	0.575 [45%-70%]	80	49	0.0589	*	*	*	>0.07	0.620 [51%-74%]	70	64	0.0387	**80**	**67**	**0.0342**
HGLZE $_{GLSZM, FBS:25HU}$	>237	0.595 [47%-72%]	90	40	0.0145	*	*	*	>583	0.577 [45%-70%]	80	50	0.0447	*	*	*
LALGLE $_{GLSZM, FBS:25HU}$	≤12	0.680 [56%-80%]	80	70	0.0007	80	52	0.1527	≤-103	0.687 [56%-80%]	60	82	0.0139	80	41	0.3502
LAHGLE $_{GLSZM, FBS:25HU}$	≤13810925	0.538[41%-66%]	90	29	0.1127	*	*	*	≤9283393	0.641[51%-76%]	70	68	0.0219	60	63	0.3819
SZN $_{GLSZM, FBS:10HU}$	>1744	0.568 [44%-69%]	80	52	0.0334	*	*	*	>1410	0.507 [38%-62%]	100	23	<0,0001	*	*	*

** Features not evaluated in the testing set because they did not fulfil the following conditions in the training set: AUC ≥ 0.60, a specificity ≥ 60% and a lack of redundancy.*

*** Bonferroni corrected p-values were significant in the testing set below 0.05 (only one feature evaluated) before harmonization and 0.017 after harmonization (3 features evaluated) respectively.*



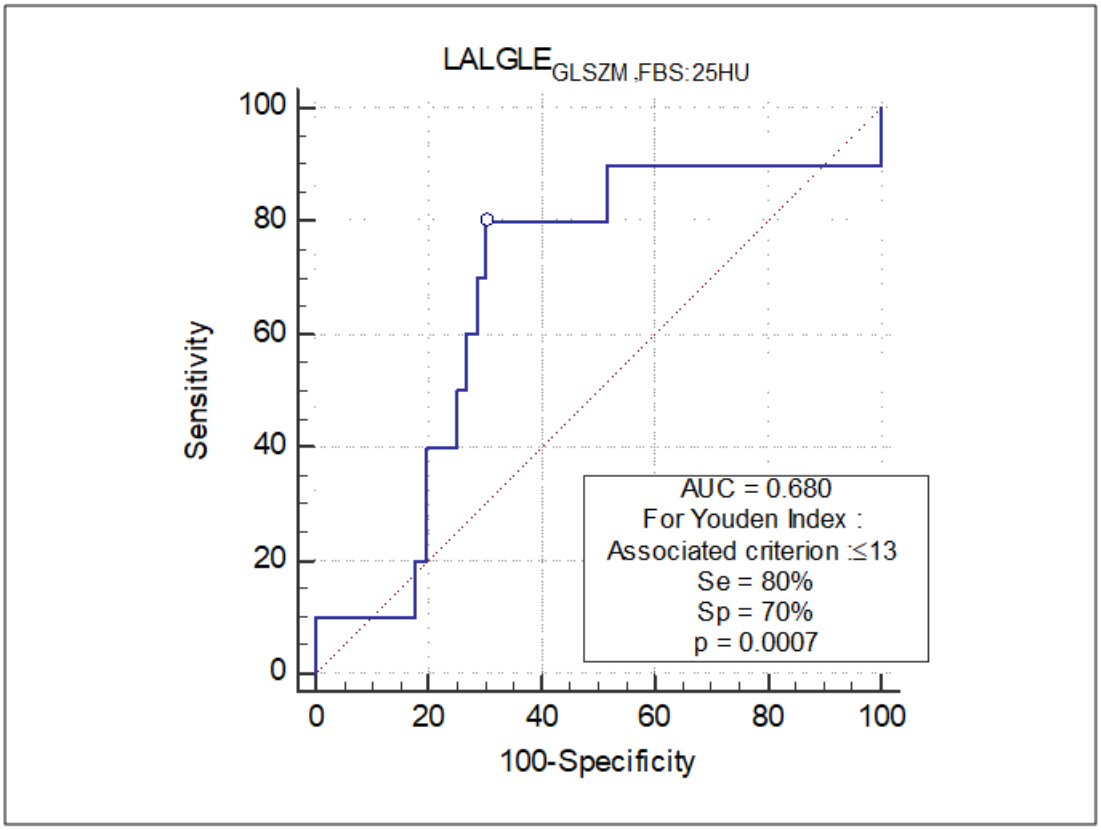
mp_14948_f1a.tif

Cluster	Samples	Non responders	Responders	Total
1	38	6	32	
2	60	9	51	
		15	83	98

mp_14948_f1b.tif

Data	Samples	Non responders	Responders	Total
Training	66	10	56	
Testing	32	5	27	
		15	83	98

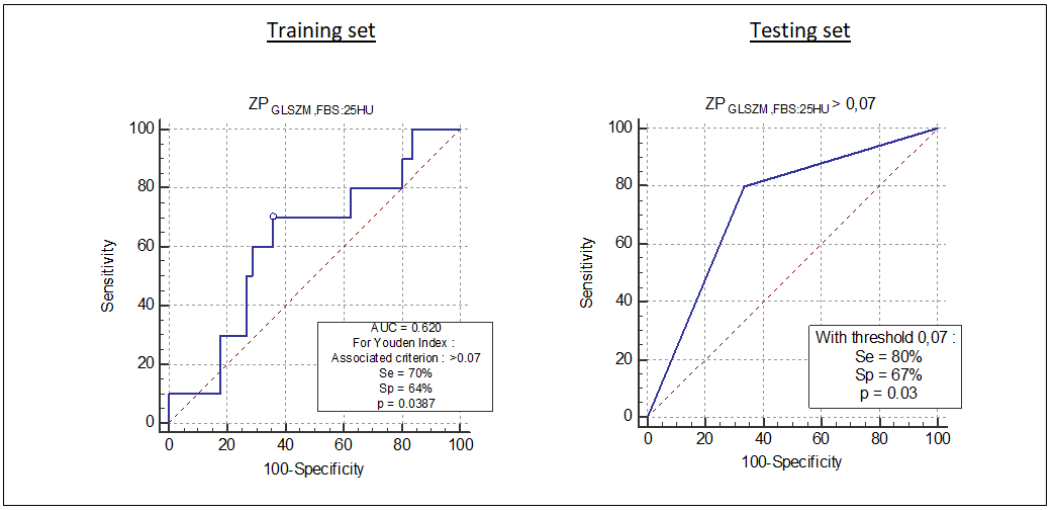
mp_14948_f1c.tif



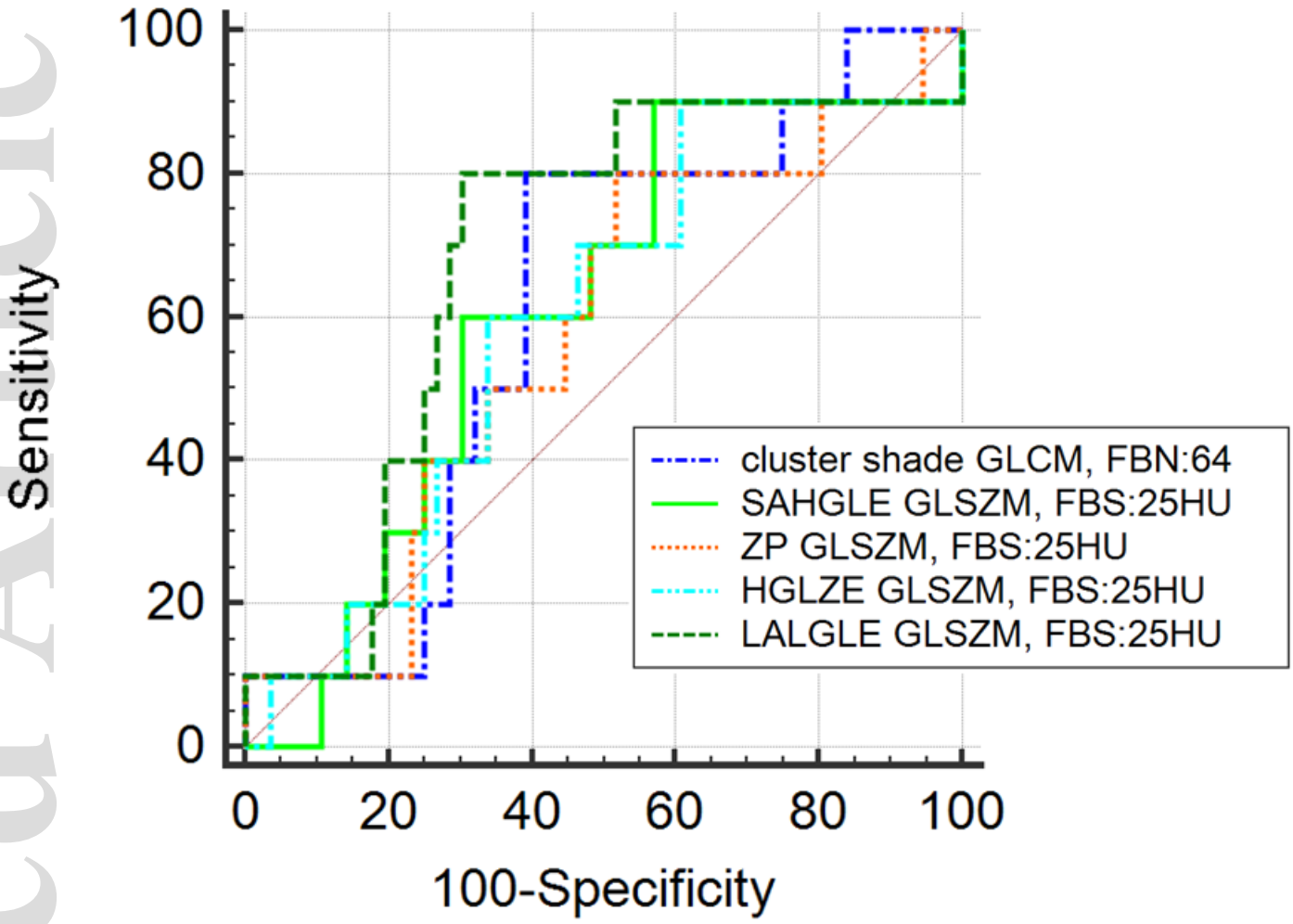
mp_14948_f2.tif

	$ZP_{GLSZM, FBS:25HU}$	$LALGLE_{GLSZM, FBS:25HU}$	$LAHGLE_{GLSZM, FBS:25HU}$
$ZP_{GLSZM, FBS:25HU}$	**1**	-0.23 [-0.45 – 0.01]	-0.22 [-0.44 – 0.02]
$LALGLE_{GLSZM, FBS:25HU}$	-0.23 [-0.45 – 0.01]	**1**	-0.14 [-0.37 – 0.11]
$LAHGLE_{GLSZM, FBS:25HU}$	-0.22 [-0.44 – 0.02]	-0.14 [-0.37 – 0.11]	**1**

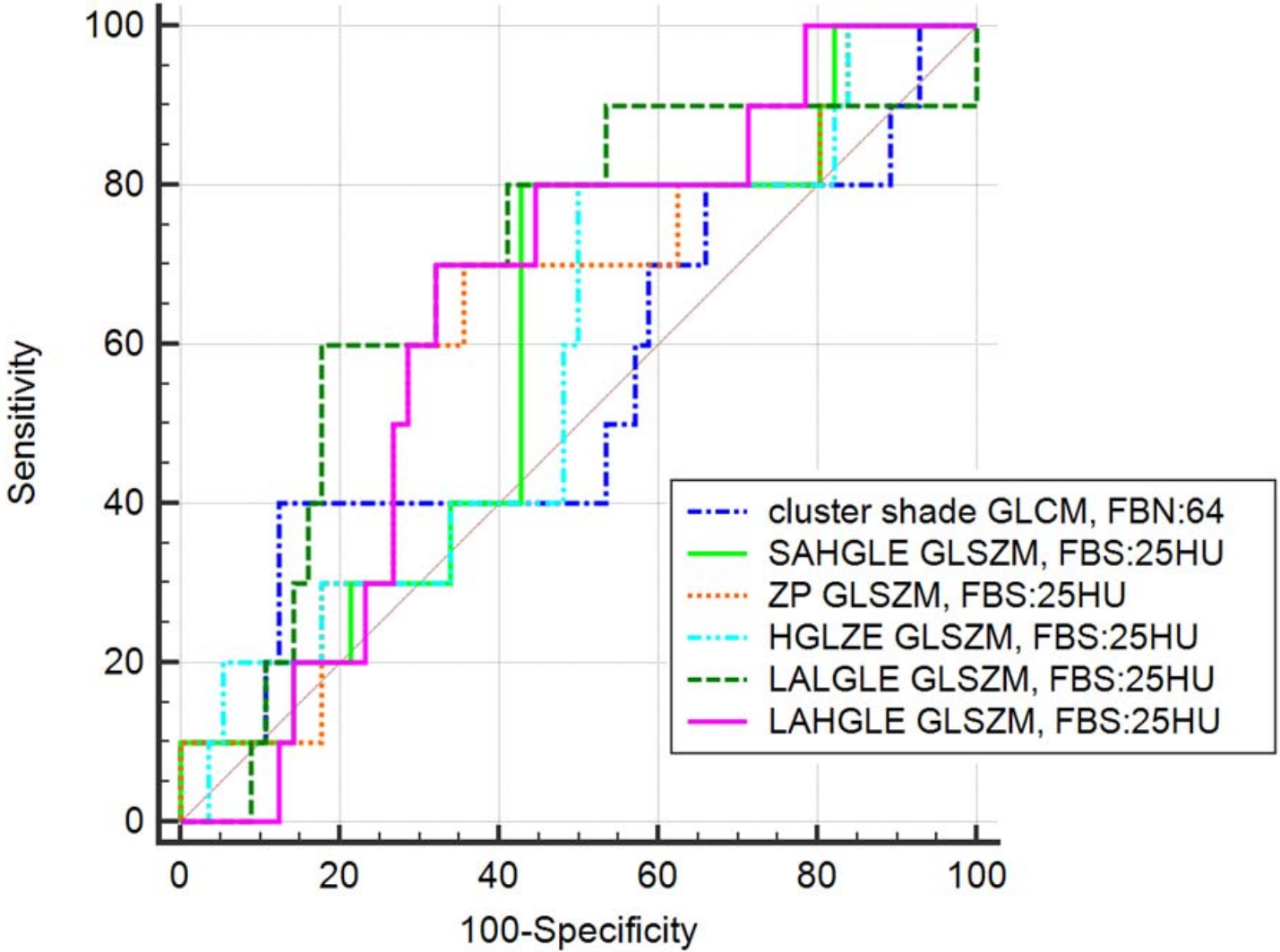
mp_14948_f3a.tif



mp_14948_f3b.tif



mp_14948_f4a.tif


100
80
60
40
20
0
Sensitivity
0
20
40
60
80
100
100-Specificity
cluster shade GLCM, FBN:64
SAHGLE GLSZM, FBS:25HU
ZP GLSZM, FBS:25HU
HGLZE GLSZM, FBS:25HU
LALGLE GLSZM, FBS:25HU
LAHGLE GLSZM, FBS:25HU