Clinical and Translational Radiation Oncology 21 (2020) 11–18

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

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

Original Research Article

Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer

Rongjie Liu [a,1], Hesham Elhalawani [b,1], Abdallah Sherif Radwan Mohamed [b,c,e], Baher Elgohari [b,f], Laurence Court [c,d], Hongtu Zhu [g], Clifton David Fuller [b,c,*]

[a] Department of Statistics, Rice University, Houston, TX 77005, USA
[b] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[c] MD Anderson UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA
[d] Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[e] Department of Clinical Oncology and Nuclear Medicine, Faculty of Medicine, University of Alexandria, Alexandria, Egypt
[f] Department of Clinical Oncology and Nuclear Medicine, Faculty of Medicine, University of Almansoura, Almansoura, Egypt
[g] Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 20 August 2019
Revised 24 November 2019
Accepted 25 November 2019
Available online 28 November 2019

2010 MSC:
00-01
99-00

Keywords:
Tumor segmentation
Oropharyngeal cancer
Stability analysis
Radiomic features

ABSTRACT

Introduction: Accurate segmentation of tumors and quantification of tumor features are important for cancer detection, diagnosis, monitoring, and planning therapeutic intervention. Due to inherent noise components in multi-parametric imaging and inter-observer and intra-observer variations, it is common that various segmentation methods may produce large segmentation errors in tumor volumes and their associated radiomic features. The purpose of this study is to carry out the stability analysis for radiomic features with respect to segmentation variation in oropharyngeal cancer (OPC).

Methods: In this study, 436 contrast-enhanced computed tomography (CT) axial images were collected from patients with OPC. In order to derive various segmentations of tumor volumes, two additional segmentations were obtained via resizing the original segmented regions of interest (ROIs) based on their geometric information on the boundary. For three ROI image groups, we calculated 109 radiomic features. Then, a logistic regression model was built to investigate the correlation between the radiomic features extracted from GTVp and the response to chemotherapy and radiation in terms of overall survival (OS). Finally, in order to evaluate the stability of each feature with respect to segmentation results, based on the prediction probabilities, we assessed the inter-rater reliability and reproducibility by calculating the intra-class correlation coefficients (ICC) and concordance correlation coefficients (CCC).

Results: Most radiomic features in this study varied a lot when the ROIs were not well segmented. For both the representation agreement and predictive agreement, the ICC and CCC were below 0.5 for all the features. We still found some robust features with relatively high ICC and CCC compared to most features. For example, 25percentile (ICC = 0.38, CCC = 0.37 in representation agreement and ICC = CCC = 0.27 in predictive agreement) is a quantile based feature, which is robust to the extremely high or low values; and Hu_1_std (ICC = 0.31, CCC = 0.31 in representation agreement) is a feature calculated based on the first Hu moment, which is invariant to the transformation of ROIs.

Conclusion: In OPC studies, the tumor segmentation variation affects the radiomic features from CT images in terms of both representation and prediction. Some features that are robust to the extreme values or invariant to the transformation of ROIs may be treated as radiomic markers to assist with OPC treatment monitoring and prognostic prediction.

1. Introduction

Medical imaging has been widely used for the management of head and neck cancers, serving as the gold standard for pre-therapy staging and post-therapy tumor control assessment [1]. As a consequence, thousands of medical imaging data are generated daily that are yet to be explored [2]. The emerging technology

* Corresponding author at: Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
E-mail address: cdfuller@mdanderson.org (C.D. Fuller).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ctro.2019.11.005
2405-6308/

of 'radiomics' (extracting quantitative imaging features from routine imaging data through a series of image processing/data-characterization algorithms) [3] further leverages existing imaging data to provide increasing degrees of predictive capacity. However, radiomic features are often affected by various techniques, including image acquisition protocols (e.g., 2D or 3D modes) [4,5], reconstruction algorithms (e.g., grid size, iteration number, and full width at half maximum) [6,7], and image preprocessing methods (e.g., tumor segmentation) [8]. Moreover, the effect of these techniques on radiomic features is still unclear. Thus, a common objective when using radiomic features is to assess the variability of radiomic features derived from different techniques.

Delineating a tumor as the volume of interest prior to radiomic feature extraction is a key rate-limiting step in quantitative imaging analysis [8]. It is common for various segmentation methods, including manual segmentation and computer-aided segmentation methods, to produce large under-segmentation and/or over-segmentation errors due to the high variability of tumor shape, size, and intensity, and the low contrast between a tumor and the surrounding tissues. For instance, manual human segmentation is time-consuming and raises concerns of inter-observer and intra-observer variability and validation. Semi-automated approaches, which are more rapid and uniform, are still being validated for applicability in different scenarios [8]. Such segmentation errors may eventually introduce large errors into the calculation of radiomic features, leading to bad predictions.

To ensure the reliability of quantitative radiomic features, accurate and robust tumor delineation is essential. Recently, more and more researches are working on the impact of variations in segmentation on radiomic features for different cancer types. For lung cancer, inter-observer variability with respect to manual versus semiautomated segmentation (3D Slicer) was studied in which semiautomated methods were found to have improved feature reproducibility in positron emission tomography (PET) image [8]. Textural feature reproducibility with respect to two different semi-automated segmentation algorithms was also studied, and homogeneity, contrast, dissimilarity, and coarseness were found to be the most reproducible features [9]. For head and neck cancer, different PET segmentation methods (manual, semiautomated, and fully automated) were compared, and more than half of the radiomic features were found to be reproducible [10]. Studies on other cancer types can be found in the recent review paper [11] and references therein. However, most existing studies focused on the PET images, while only one multi-center study was designed for CT image [12]. Thus, it is of great importance to pay more attention to the impact of variations in CT image segmentation.

To that end, our team carried out stability analysis of standard radiomic features with respect to manual segmentation variability based on the contrast-enhanced computed tomography (CT) axial images of 436 patients with oropharyngeal cancer (OPC). Specifically, we created three sets of tumor segmentation results, representing manual segmentation results, under-segmentation results, and over-segmentation results. We assessed the inter-rater reliability and reproducibility of the three segmentation results via calculating the intra-class correlation coefficient (ICC) and concordance correlation coefficient (CCC), which are commonly used in existing literature [11]. To test reproducibility, we assessed the correlation to treatment-related outcomes as a function of the radiomic features.

2. Material and methods

2.1. Patient demographics and clinical end points

The 436 patients with OPC treated with curative-intent intensity-modulated radiation therapy (IMRT) at The University of Texas MD Anderson Cancer Center were drawn from a larger oropharynx cohort between the years 2005 and 2012. Patients were retrieved from an internal University of Texas MD Anderson Cancer Center database after getting approved by the University of Texas MD Anderson Cancer Center Institutional review board (IRB). All methods for this study were performed in accordance with the University of Texas MD Anderson Cancer Center IRB guidelines and regulations. Being an HIPAA-compliant retrospective study waived the prerequisite for informed consent. The records of all the 436 patients were thoroughly screened for specific demographic data, disease characteristics, treatment details and outcomes [13]. In particular, the patients' demographics data included: gender, age at diagnosis and race. Disease characteristics encompassed: tumor laterality and oropharynx subsite of origin. Furthermore, TNM (Tumor, node and metastases) classification was also provided, where T category described the original (primary) tumor, as regard its size and extent, per the American Joint Committee on Caner (AJCC) and Union for International Cancer Control (UICC) cancer staging system, 7th edition. Similarly, N category described whether or not the cancer has reached nearby lymph nodes, per the AJCC and UICC cancer staging system, 7th edition, along with the corresponding AJCC stage. Also, individual patient's vital status was dichotomously reported as '1 = alive' or '0 = dead', as an indicator for overall survival (OS) status. More details about the patient demographics and clinical information can be found in the work of Elhalawani et al [13].

2.2. Image acquisition

Contrast-enhanced CT images were performed independently in the course of pre-treatment diagnostic work-up according to institutional protocol. The contrast-enhanced CT images were restored from the patients' electronic medical records, with a section thickness of 1–5 mm (median: 1.25 mm in 84.7% of the cases) and an X-ray tube current of 100–584 mA (220 mA for 68.1% of the patients) at 100–154 kVp (120 kVp for 66% of the patients). Most of the CT scans (92%) were obtained using GE Medical Systems scanners, LightSpeed16 (55.2%) and LightSpeed VCT (27.4%) models. The display field of view was 25 cm; axial images were acquired by using a matrix of 512×512 pixels and reconstructed with a voxel size of 0.048828 cm $\times$ 0.048828 cm along the x-axis and y-axis. Forty-four patients had CT scans with a slice thickness (Z-dimension) that was not equal to 1.25 mm (range 0.5 to 5 mm). One hundred and twenty milliliters of contrast material were injected at a rate of 3 ml sec^{-1}, followed by scanning after a 90-s delay [13].

2.3. Image preprocessing

Some preprocessing steps, including smoothing, normalization, and resampling, were applied on the raw CT images. In order to implement these steps, an open infrastructure software platform, named "IBEX" [14], were considered here. Specifically, the Gaussian smoothing method and histogram matching method [15] were selected for the smoothing step and normalization step, respectively. For those forty-four CT scans with a slice thickness not equal to 1.25 mm, the UpDown-Sampling method was adopted to make all the images with the consistent slice thickness, i.e., 1.25 mm.

2.4. Manual segmentation of regions of interest (ROIs)

We performed manual segmentation by using commercial treatment planning software VelocityAI 3.0.1. Gross tumor volumes (GTVs) for the primary tumor (GTVp) constituted our regions of interest (ROIs) for this project. The segmentation result of one randomly selected patient is shown in Fig. 1. Tumor volumes were

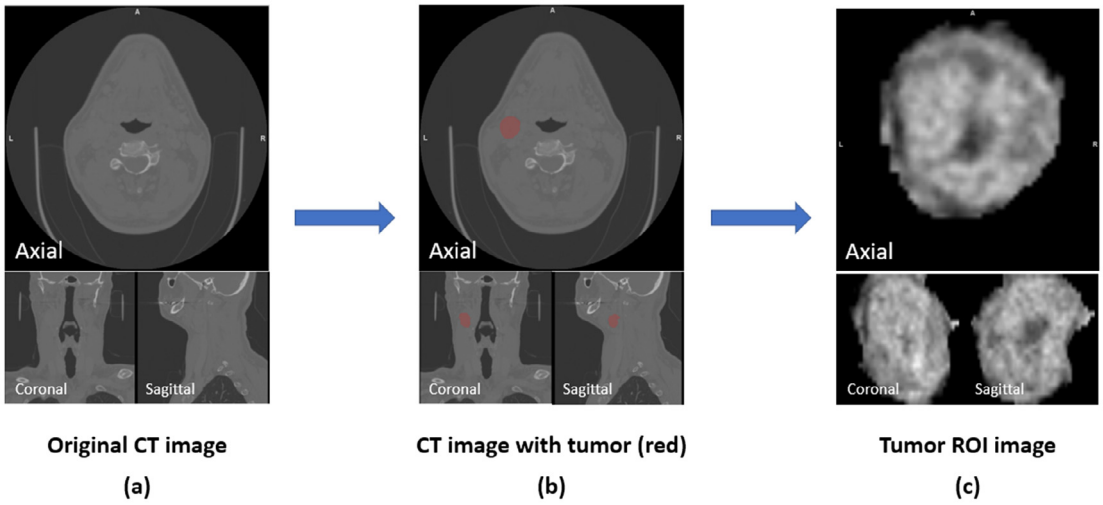

Fig. 1. Illustration of manual segmentation result. (a) original CT image; (b) CT image with tumor highlighted in red; (c) segmented tumor ROI image.

manually segmented on each individual patient's diagnostic contrast-enhanced CT axial images and simulation CT images by the collaborators independently. They were blinded to relevant clinical meta-data, and a radiation oncologist (H.E.) revised the segmentation according to the regulations that we followed for previous projects [16]. The segmentation process was governed by the guidelines of the International Commission on Radiation Units and Measurements, report 83. Segmentation primarily relied on the findings from physical examination, fiberoptic nasopharyngolaryngoscopy and imaging studies.

2.5. Under-segmentation and over-segmentation ROIs

To understand the effects of different segmentation results on radiomic features, the original manually segmented ROIs were resized based on their geometric information, i.e., normals, on the surface. For simplicity, instead of considering 3D surface, here we focused on the middle slice along the Z-axis, which contains the biggest area and the most information of ROI. In particular, the normal to a 2D closed curve at a point **P** is a vector that is perpendicular to the line that is tangential to that closed curve at **P**. The normalized outer-pointing normal, $\mathbf{N_P}$, of point **P** on the boundary was calculated by using the corresponding neighboring points. In order to guarantee the uniqueness of normal vectors, only the normalized outer-pointing normal was considered here.

For a given interior point **Q** inside the ROI, the distance between points **P** and **Q**, denoted as d_{PQ}, was calculated. The aim of introducing point **Q** is to address the potential issue when the closed ROI curve is with concave shape. Assuming that the resize scale factor is **a**, the point $\mathbf{P_{new}}$ on the curve of the resized ROI was calculated as

$$\mathbf{P_{new}} = \mathbf{P} + (\mathbf{a} - \mathbf{1})\mathbf{d_{PQ}}\mathbf{N_P}. \quad (1)$$

According to the formula (1), it can be found that, the distance between new point **P_new** and **P** strongly depends on the distance between **P** and the chosen point **Q**. Then, for the ROI in concave shape, the point **Q** is useful since it can be chosen to make sure the new generated ROI boundary is a simple closed curve without self-crossing.

The pipeline to construct modified segmented ROIs is illustrated in Fig. 2. The over-segmentation and under-segmentation groups were created by respectively setting different resize scale factors. We tried multiply scale ratios, like 0.7, 0.8, 0.9 & 1.1, 1.2, 1.3. However, we found that the down-stream analysis results don't vary too much across different scale ratios for neither under nor over segmentations. Here we only conducted all the down stream analysis for two resize scale factors: a = 1.2 and a = 0.8.

2.6. Radiomic features

For each ROI of the three segmentation groups, we used an internally developed application to import the ROI data file and then calculated 109 radiomic features, including first-order features, second-order textural features, shape descriptors, and wavelet-based features from each ROI. Detailed features can be found in Tables A1–A4 (Supplementary materials). For a total of 109 image features, 9 first-order features (extracted from image intensity statistics) were calculated directly from volume intensity histograms; 13 second order gray-level co-occurrence matrix (GLCM) features, originally described by Haralick et al. [17], were implemented based on the gray-level matrix metric; 45 moment-based shape descriptors were calculated by the Hu-moments [18] and Zernike-moments [19] for each 2D slice along the Z-direction. The number of moment features shown in Table A3 (Supplementary materials) is 6; noticing that some first-order operators (e.g., max, min, and std.) were applied on each moment across all the slices; 42 wavelet-based features were calculated by local binary pattern (LBP) and threshold adjacency statistics (TAS) [20] For example, 5 first-order features were calculated from the LBP map [21], 3 first-order features were calculated from the local feature maps derived based on TAS [22], and 9 first-order features were calculated from the wavelet-based local feature maps. The detailed formulas for first-order features, shape based features, GLCM features, and wavelet features can be found in existing literature[23,24].

2.7. Prediction based on radiomic features

Our first goal was to identify a set of top radiomic features in the prediction of OS of patients with OPC among all 109 imaging features using receiver operating characteristic (ROC) analysis across the three segmented ROI groups. We ranked these features according to their area under the ROC curve (AUC) values. For each of the three ROI image groups, the top 10 ranked features were cross-validated by the following scheme of leave 1/3-out cross-validation:

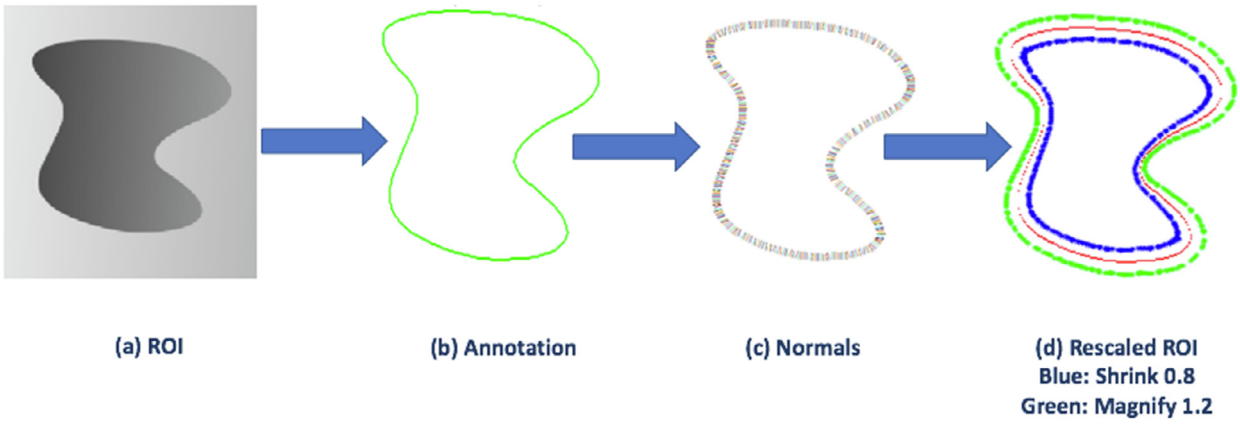

Fig. 2. Illustration of manual segmentation rescaling. (a) original ROI; (b) annotation; (c) normalized outer-pointing normals; (d) rescaled ROIs (blue: over-segmentation ROI with scale factor 0.8; green: under-segmentation ROI with scale factor 1.2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1. Split the data into a training set (2/3 of the data) and a test set (1/3 of the data) of the data.
2. Fit the logistic regression model using the training set and apply the model to the test data in order to calculate the AUCs.
3. Repeat this process 1000 times and average out the obtained AUCs.

The validation set is not needed here because there is no any tuning parameters when we apply the logistic regression model on each feature for classification. The aim of repeating the splitting process for 1000 times is to average the obtained AUC for each feature, which can avoid the instability of one-time data random splitting. We also set the random seeds in the R code for reproducibility.

2.8. Stability analysis

Our primary goal was to evaluate the stability of each radiomic feature with respect to the segmentation results. We investigated two aspects of stability: feature representation and prediction. For each aspect, we assessed the inter-rater reliability and reproducibility via calculating the ICC [25] and CCC [26].

The ICC is a statistical measure that ranges between −1 and 1, indicating null and perfect reproducibility, respectively. ICC values for the intra-observer segmentations were obtained from one-way analysis of variance. Similarly, the CCC ranges between −1 and 1. A value of 1 corresponds to perfect agreement; a value of −1 corresponds to perfect negative agreement; and a value of 0 corresponds to no agreement. In terms of the representation of each feature, the ICC and CCC were calculated directly based on the feature values from each pair of the ROI groups. Whereas in terms of feature-based prediction, the ICC and CCC were calculated based on the predictive probabilities derived from the logistic regression model. In particular, there were three comparisons for each feature: original ROI group vs. over-segmentation ROI group; original ROI group vs. under-segmentation ROI group; and over-segmentation ROI group vs. under-segmentation ROI group. For each comparison, the point estimate and lower and upper bounds of the confidence interval of both ICC and CCC were calculated by using the R package agRee (various methods for measuring agreement).

3. Results

3.1. Patients' disease information

The information regarding the patients' disease is included in Table A5 (supplementary materials). Of note, for this cohort, the median age was 58 years, 86% (n = 374) were male, and 90.8% (n = 395) were Caucasian. There was a relatively even distribution of patients when considering smoking status, where 21.6%, 37%, and 41.4% were current, former, or never smokers, respectively. Regarding HPV status, 52.4% (n = 228) were HPV+, 10.8% (n = 47) were HPV-, while the rest (36.8%) had 'unknown' HPV status coded in the electronic medical records. Primary tumors most commonly originated at the base of the tongue or the tonsillar region (52.6% and 39.5%, respectively). For the entire patient cohort, tumors were most commonly locally advanced neoplasms (95.5%).

3.2. Top 10 radiomic features in term of AUC

For each ROI segmentation group, the top 10 radiomic features with highest AUCs are listed in Table 1. First, in the original ROI group, there are two features with AUCs above 0.7, including GLCM_sum_var, and GLCM_contrast. There is no features with this level of AUC in neither under- nor over-segmentation ROI groups. This may indicate that the original ROI segmentation group outperforms the other two segmentation groups in OS prediction. Second, the average AUCs for the top 10 features in the over-segmentation ROI group are lower than those in the under-segmentation ROI group. This result may be reasonable since the magnification of these over-segmentation ROIs may increase heterogeneity within the ROI. It might increase prediction errors when a specific feature might be sensitive to such heterogeneity. Third, the GLCM-based features (e.g., GLCM_sum_var) performed much better in the original ROI group than in the under- and over-segmentation groups. This may indicate that these traditional features are sensitive to segmentation results in terms of prediction of OS.

3.3. Top radiomic features in terms of ICC & CCC

Table 2 presents the ICC and CCC values of the top 5 radiomic features with highest representation agreement across all three ROI groups. The lower and upper bounds of the 90% confidence intervals of ICC and CCC are presented as well. In order to show the prediction performance for all the top features, the corresponding averaged AUCs are also presented in Table 2. In comparing the original ROI group with each resized ROI group, the ICC and CCC associated with the top feature are almost the same and both are above 0.3. Moreover, both the ICC and CCC are close to 0 for most features. Considering the comparison of the original ROI group and the under-segmentation ROI group, the top feature is Hu_1_std, which is a feature based on the Hu Moment and invariant to the transformation of ROIs. Considering the comparison between the

Table 1
Top 10 features with highest AUCs in prediction of OS.

	Original ROI		Under-segmentation ROI		Over-segmentation ROI	
	Feature	AUC	Feature	AUC	Feature	AUC
1	GLCM_contrast	0.71	Zernike_3_std	0.69	Zernike_8_min	0.67
2	GLCM_sum_var	0.70	Hu_5_std	0.68	Zernike_5_max	0.67
3	Entropy	0.70	TAS_5_max	0.67	TAS_4_max	0.65
4	Zernike_7_max	0.68	Zernike_4_max	0.65	Hu_3_std	0.65
5	GLCM_diff_var	0.66	Hu_2_std	0.65	DWT_25percentile	0.65
6	Std	0.66	DWT_75percentile	0.65	Zernike_3_std	0.65
7	GLCM_var	0.66	LBP_min	0.65	25percentile	0.64
8	GLCM_entropy	0.65	25percentile	0.62	Zernike_4_std	0.63
9	TAS_2_max	0.65	Min	0.62	Hu_1_max	0.62
10	25percentile	0.65	Zernike_8_std	0.62	75percentile	0.61

Table 2
Top 5 features with highest representation agreement when comparing each pair of ROI groups. Original = original segmentation ROI group; Under = under-segmentation ROI group; Over = over-segmentation ROI group; ICC = intra-class correlation coefficient; C.I. = 90% confidence interval; CCC = concordance correlation coefficient; AUC = averaged AUCs of features in each of the two ROI groups respectively.

Original vs Under	Feature	ICC	C.I.	AUC	Feature	CCC	C.I.	AUC
	Hu_1_std	0.31	[0.24, 0.38]	[0.58, 0.62]	Hu_1_std	0.31	[0.24, 0.38]	[0.58, 0.62]
	TAS_6_std	0.12	[0.04, 0.20]	[0.64, 0.58]	TAS_6_std	0.13	[0.06, 0.21]	[0.64, 0.58]
	TAS_5_std	0.12	[0.04, 0.20]	[0.60, 0.59]	TAS_5_std	0.13	[0.05, 0.20]	[0.60, 0.59]
	TAS_4_std	0.09	[0.01, 0.17]	[0.59, 0.57]	TAS_4_std	0.09	[0.01, 0.16]	[0.59, 0.57]
	TAS_2_std	0.07	[−0.01, 0.15]	[0.58, 0.58]	TAS_2_std	0.07	[−0.01, 0.15]	[0.58, 0.58]
Original vs Over	Feature	ICC	C.I.	AUC	Feature	CCC	C.I.	AUC
	25percentile	0.38	[0.31, 0.44]	[0.65, 0.64]	25percentile	0.37	[0.30, 0.44]	[0.65, 0.64]
	TAS_1_min	0.13	[0.05, 0.21]	[0.64, 0.56]	max	0.07	[0.02, 0.13]	[0.65, 0.59]
	TAS_2_min	0.12	[0.05, 0.20]	[0.58, 0.58]	entropy	0.07	[−0.01, 0.15]	[0.69, 0.58]
	TAS_3_min	0.12	[0.04, 0.20]	[0.56, 0.59]	kurtosis	0.06	[−0.03, 0.13]	[0.63, 0.58]
	TAS_6_min	0.12	[0.04, 0.20]	[0.58, 0.56]	median	0.04	[−0.03, 0.12]	[0.64, 0.56]
Under vs Over	Feature	ICC	C.I.	AUC	Feature	CCC	C.I.	AUC
	TAS_1_min	0.06	[−0.01, 0.14]	[0.59, 0.56]	max	0.02	[−0.02, 0.07]	[0.60, 0.59]
	TAS_1_std	0.04	[−0.04, 0.12]	[0.57, 0.57]	entropy	0.02	[−0.06, 0.09]	[0.60, 0.58]
	median	0.02	[−0.06, 0.10]	[0.54, 0.56]	median	0.02	[−0.06, 0.09]	[0.54, 0.56]
	LBP_median	0.01	[−0.07, 0.09]	[0.54, 0.54]	std	0.01	[−0.07, 0.09]	[0.60, 0.58]
	std	0.01	[−0.07, 0.09]	[0.60, 0.58]	Hu_1_max	0.00	[−0.00, 0.00]	[0.60, 0.62]

original ROI group and the over-segmentation ROI group, the top feature is 25percentile, which is robust to the extremely high and low values. In addition, in a comparison of the under- and over-segmentation ROI groups, the ICC and CCC values associated with the top feature are both below 0.1, which means that all of the features show very low representation agreement between the two resized ROI groups. Fig. 3 confirms this finding based on the Bland–Altman plot.

The ICC and CCC values of the top 5 features with highest predictive agreement in comparing each pair of ROI groups are listed in Table 3, along with the lower and upper bounds of the 90% confidence intervals of ICC and CCC. Here the corresponding averaged AUCs of all the top features are presented as well. In comparing the original ROI group with the under- and over-segmentation ROI groups, the top feature is the same, i.e., 25percentile, with the ICC and CCC associated with the top feature being the same, 0.27. Thus, we may have the following findings: (i) First, 25percentile is very robust to heterogeneity when comparing the feature representation of the original ROI group and the over-segmentation ROI group; and (ii) Second, 25percentile is among the top 10 features with highest AUCs when predicting OS from all the three ROI groups [27]. Therefore, feature 25percentile can be treated as a potential radiomic marker with robustness of segmentation variability. Another interesting finding is that GLCM-based features (e.g., GLCM_ang_2m) have low ICC and CCC (<0.3) when comparing the original ROI group and each resized ROI group; whereas they have high ICC and CCC (>0.9) when comparing the two resized ROI groups. In spite of the high ICC and CCC, these GLCM-based features were found to show poor performance when predicting OS according to their AUCs. Specifically, for both under- and over-segmentation groups, all the prediction probabilities are above 0.6 and most of them are around 0.8, which leads to a serious false positive error. Thus, in the comparison between under- and over-segmentation groups, all the top features are random predictors although they have high predictive agreement. Fig. 4 confirms this finding based on the Bland–Altman plot and scatter plot.

4. Discussion

This study carried out stability analysis for various radiomic features with respect to segmentation results based on the contrast-enhanced CT axial images of 436 patients with OPC. The segmentation variability was illustrated via resizing the original segmented ROIs and constructing under- and over-segmentation ROIs. For the three segmentation ROI groups, 109 radiomic features were calculated, and a logistic regression model was built to find the top features with high AUCs in predicting OS. ICC and CCC were adopted to assess the representation and predictive agreement when comparing each pair of segmentation ROI groups.

According to the results, it can be induced that the radiomic feature values vary a lot when the ROIs are not well segmented. Specifically, in comparing the original ROI group and the resized ROI group, both the ICC and CCC were below 0.5 for all the features (see Table 2 and Fig. 3), and close to 0 for most of the features. Nevertheless, we still found some robust features with relatively high

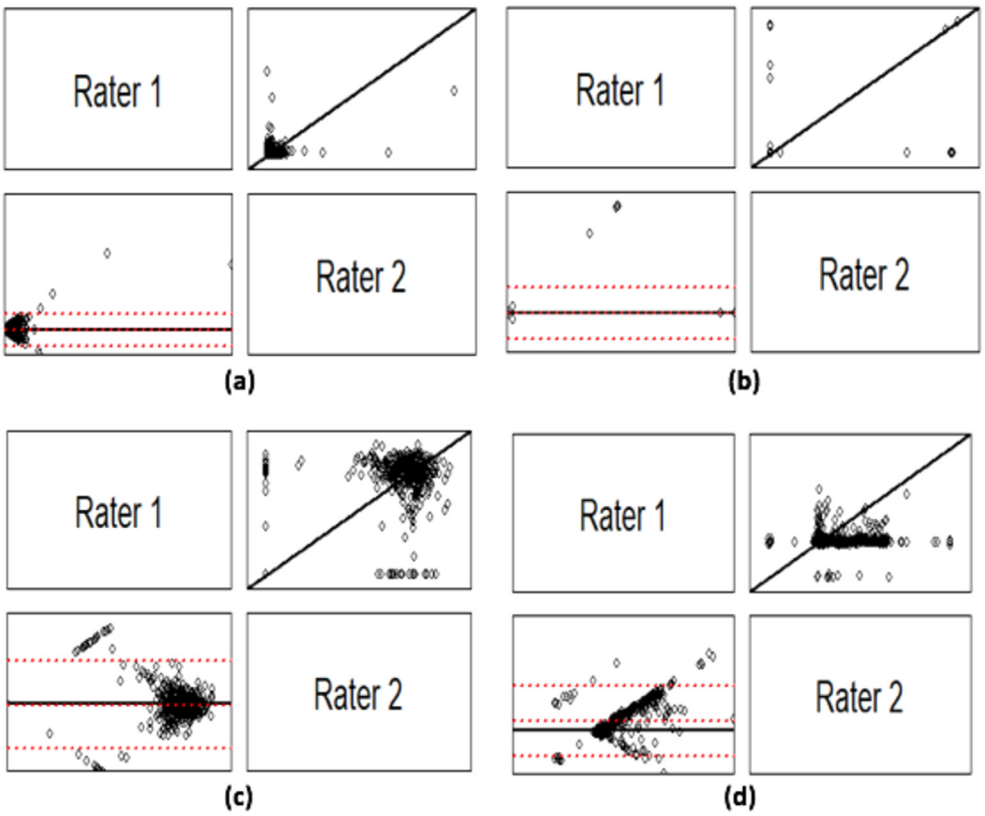

Fig. 3. Bland-Altman plot (bottom left) and scatter plot (top right) for the top feature with highest representation agreement when comparing each pair of ROI groups. (a) Hu_1_std (ICC = 0.31, CCC = 0.31) in original vs. under-segmentation; (b) 25percentile (ICC = 0.38, CCC = 0.37) in original vs. over-segmentation; (c) TAS_1_min (ICC = 0.06) in under-segmentation vs. over-segmentation; (d) max (CCC = 0.02) in under-segmentation vs. over-segmentation.

Table 3
Top 5 features with highest representation agreement when comparing each pair of ROI groups. Original = original segmentation ROI group; Under = under-segmentation ROI group; Over = over-segmentation ROI group; ICC = intra-class correlation coefficient; C.I. = 90% confidence interval; CCC = concordance correlation coefficient; AUC = averaged AUCs of features in each of the two ROI groups respectively

Original vs Under	Feature	ICC	C.I.	AUC	Feature	CCC	C.I.	AUC
	25percentile	0.27	[0.19, 0.34]	[0.65, 0.62]	25percentile	0.27	[0.22, 0.32]	[0.65, 0.62]
	TAS_6_max	0.14	[−0.12, 0.24]	[0.57, 0.57]	GLCM_sum_var	0.02	[−0.01, 0.06]	[0.70, 0.61]
	TAS_9_std	0.08	[−0.11, 0.15]	[0.63, 0.54]	GLCM_var	0.01	[−0.02, 0.05]	[0.66, 0.62]
	Zernike_3_std	0.05	[−0.08, 0.08]	[0.57, 0.69]	kurtosis	0.01	[−0.07, 0.09]	[0.63, 0.58]
	Zernike_6_max	0.03	[−0.08, 0.09]	[0.56, 0.57]	Hu_2_std	0.01	[−0.06, 0.07]	[0.57, 0.65]
Original vs Over	Feature	ICC	C.I.	AUC	Feature	CCC	C.I.	AUC
	25percentile	0.27	[0.19, 0.34]	[0.65, 0.64]	25percentile	0.27	[0.21, 0.32]	[0.65, 0.64]
	DWT_min	0.16	[−0.09, 0.17]	[0.61, 0.54]	Hu_1_max	0.01	[−0.07, 0.08]	[0.58, 0.62]
	TAS_8_max	0.14	[−0.04, 0.16]	[0.63, 0.58]	GLCM_corr	0.01	[−0.03, 0.04]	[0.63, 0.59]
	Zernike_8_max	0.11	[−0.03, 0.19]	[0.64, 0.58]	median	0.00	[−0.03, 0.04]	[0.64, 0.56]
	Zernike_6_min	0.07	[−0.01, 0.15]	[0.62, 0.58]	kurtosis	0.00	[−0.01, 0.01]	[0.63, 0.58]
Under vs Over	Feature	ICC	C.I.	AUC	Feature	CCC	C.I.	AUC
	GLCM_ang_2m	0.97	[0.97, 0.98]	[0.53, 0.54]	GLCM_ang_2m	0.97	[0.97, 0.98]	[0.53, 0.54]
	TAS_8_min	0.97	[0.97, 0.98]	[0.57, 0.55]	GLCM_inv_diff_m	0.95	[0.94, 0.96]	[0.60, 0.60]
	LBP_25percentile	0.96	[0.89, 0.92]	[0.60, 0.60]	GLCM_diff_var	0.95	[0.94, 0.95]	[0.60, 0.60]
	LBP_75percentile	0.96	[0.82, 0.96]	[0.56, 0.54]	GLCM_sum_entropy	0.93	[0.92, 0.94]	[0.59, 0.58]
	GLCM_inv_diff_m	0.95	[0.94, 0.96]	[0.60, 0.60]	GLCM_entropy	0.93	[0.92, 0.94]	[0.60, 0.59]

ICC and CCC compared to most features. For example, 25percentile (ICC = 0.38, CCC = 0.37) is a quantile based feature, which is robust to the extremely high or low values; and Hu_1_std (ICC = 0.31, CCC = 0.31) is a feature calculated based on the first Hu moments, which is invariant to the transformation of ROIs. Thus, in order to avoid the influence of under- and over-segmentation errors, in the future it will be helpful to seek or develop some features that are either robust to the extreme values or invariant to the transformation of ROIs.

Besides the feature representation agreement, it is also of great importance to discuss the link with segmentation variability and predictive accuracy. According to the results, it was found that the prediction performance decreased when the ROIs were either under- or over-segmented in terms of averaged AUCs. For example,

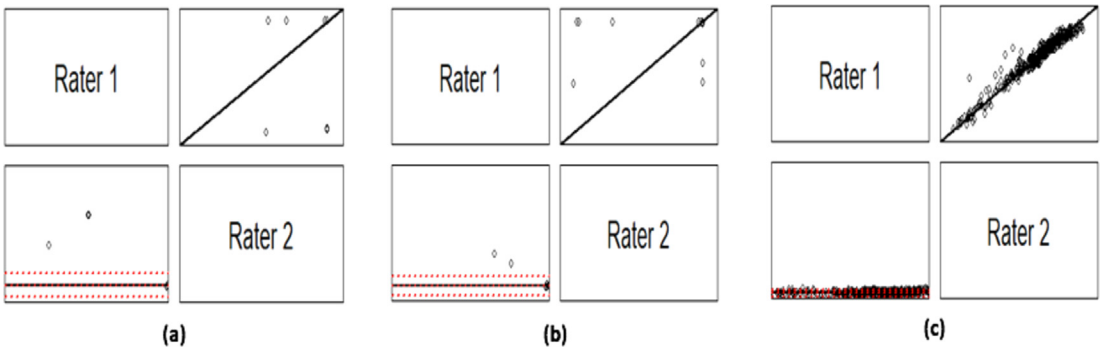

Fig. 4. Bland-Altman plot (bottom left) and scatter plot (top right) for top feature with highest predictive agreement when comparing each pair of ROI groups. (a) 25percentile (ICC = 0.27, CCC = 0.27) in original vs. under-segmentation; (b) 25percentile (ICC = 0.27, CCC = 0.27) in original vs. over-segmentation; (c) GLCM_ang_2m (ICC = 0.97, CCC = 0.97) in under-segmentation vs. over-segmentation.

the averaged AUCs of traditional GLCM-based features could achieve around 0.7 for the original ROI group while only 0.6 or even lower for the other two ROI groups (see Table 1). Although these GLCM-based features had really high predictive agreement in comparing under- and over-segmentation ROI groups, they could only be treated as random features since there was serious false positive error in the prediction probabilities. In addition, it was also found that the prediction performance for the under-segmentation ROI group outperformed that for the over-segmentation ROI group, which could be caused by the additional heterogeneity introduced by the over-segmented ROIs. Thus, it will be meaningful to derive some features that are not sensitive to the heterogeneity.

Finally, we still found some reasonable features that had relatively high predictive agreement in terms of ICC and CCC when comparing the original ROI group with other two groups. For example, 25percentile was the top one feature with highest ICC and CCC in both two comparisons (see Table 3). It also showed great consistency across different ROI groups in terms of averaged AUCs (see Table 3). Therefore, the feature 25percentile, which is robust to segmentation variability in terms of both representation and prediction, may be treated as a potential radiomic marker to assist with OPC treatment monitoring and prognostic prediction.

Some future research directions are considered here. First, although we have 3D ROI data, this paper conducted all the comparisons and analysis only on the selected 2D slice. In the future, it will be interesting to see all the stability analysis based on the entire 3D ROIs. Second, besides the segmentation step, some other steps, e.g., normalization, in the preprocessing pipeline are also have potential influences on the feature's representation and prediction accuracy. So it will be also interesting to conduct all the stability analysis on those steps. Third, as we mentioned in the image acquisition section, two different scanner models (LightSpeed16 and LightSpeed VCT models in GE medical system scanner) were considered. For some other image modalities, e.g. PET image, the stability analysis of radiomic features with respect to image acquisition protocol variation has been investigated [4]. Thus, it is of great importance to conduct the stability analysis on CT images in oropharyngeal cancer studies.

Author contributions statement

R.L., H.E., C.F. and H.Z. were responsible for the conception and design of the study; H.E., A.M., B.E., L.C., and C.F. conducted the experiments; R.L. performed data-analysis; R.L. and H.E. drafted the manuscript; C.F. and H.Z. provided critical revision; all authors approved the final draft of the manuscript for submission.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.ctro.2019.11.005.

References

[1] Sumi M, Kimura Y, Sumi T, Nakamura T. Diagnostic performance of mri relative to ct for metastatic nodes of head and neck squamous cell carcinomas. J Magn Reson Imag 2007;26(6):1626–33.
[2] Campbell S, Mohamed A, Heukelom J, Awan M, Garden A, Gunn G, Rosenthal D, Fuller C. Primary tumor and nodal regression rates: the prognostic value of volumetric image guided radiation therapy for head and neck cancer. Int J Rad Oncol Biol Phys 2016;94(4):922.
[3] Bourgier C, Colinge J, Ailleres N, Fenoglietto P, Brengues M, Pelegrin A, Azria D. Radiomics: definition and clinical development. Cancer Radiother 2015;19(6–7):532–7.
[4] Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in fdg pet images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49(7):1012–6.
[5] Zhao B, Tan Y, Tsai W-Y, Qi J, Xie C, Lu L, Schwartz LH. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific Rep 2016;6:23428.
[6] Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, Tham I, Townsend D. Impact of image reconstruction settings on texture features in 18f-fdg pet. J Nucl Med 2015;56(11):1667–73.
[7] Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of image features computed from conventional and respiratory-gated pet/ct images of lung cancer. Transl Oncol 2015;8(6):524–34.
[8] Parmar C, Velazquez ER, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, Mitra S, Shankar BU, Kikinis R, Haibe-Kains B, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9(7):e102107.
[9] Orlhac F, Soussan M, Maisonobe J-A, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18f-fdg pet: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med 2014;55(3):414–22.
[10] Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, Chen W. Robustness of radiomic features in [11 c] choline and [18 f] fdg pet/ct imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization. Mol Imag Biol 2016;18(6):935–45.
[11] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Rad Oncol Biol Phys 2018.
[12] Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, Echegaray S, Rubin D, McNitt-Gray M, Lo P, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. Tomography 2016;4:430.
[13] Elhalawani H, Mohamed AS, White AL, Zafereo J, Wong AJ, Berends JE, AboHashem S, Williams B, Aymard JM, Kanwar A, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Scientific Data 2017;4:170077.

[14] Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. Ibex: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys 2015;42(3):1341–53.
[15] Gonzales RC, Woods RE. Digital Image Process 2002.
[16] Mohamed AS, Rosenthal DI, Awan MJ, Garden AS, Kocak-Uzel E, Belal AM, El-Gowily AG, Phan J, Beadle BM, Gunn GB, et al. Methodology for analysis and reporting patterns of failure in the era of imrt: head and neck cancer applications. Rad Oncol 2016;11(1):95.
[17] Haralick RM, Shanmugam K, et al. Textural features for image classification. IEEE Trans Systems Man Cybern 1973;6:610–21.
[18] Hu M-K. Visual pattern recognition by moment invariants. IRE Trans Inf Theory 1962;8(2):179–87.
[19] Teague MR. Image analysis via the general theory of moments. JOSA 1980;70 (8):920–30.
[20] Valkonen M, Kartasalo K, Liimatainen K, Nykter M, Latonen L, Ruusuvuori P. Metastasis detection from whole slide images using local features and random forests. Cytometry Part A 2017;91(6):555–65.
[21] He D-C, Wang L. Texture unit, texture spectrum, and texture analysis. IEEE Trans Geosci Remote Sens 1990;28(4):509–12.
[22] Hamilton NA, Pantelic RS, Hanson K, Teasdale RD. Fast automated cell phenotype image classification. BMC Bioinf 2007;8(1):110.
[23] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.
[24] Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJ, Khaouam N, Nguyen-Tan PF, Wang C-S, Sultanem K, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific Rep 2017;7 (1):10117.
[25] Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. Biometrics 1980:19–25.
[26] Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989:255–68.
[27] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems. J Mach Learn Res 2014;15(1):3133–81.