Technical note

Robustness and reproducibility of radiomics in T2 weighted images from magnetic resonance image guided linear accelerator in a phantom study

Mengdi Sun a, e, Ahmad Baiyasi b, Xuechun Liu c, d, Xihua Shi d, Xu Li d, Jian Zhu d, Yong Yin d, Jiani Hu b, Zhenjiang Li a, b, c, Baosheng Li a, b, d, e

Department of Radiology, Wayne State University, Detroit, United States
Department of Radiation Oncology Physics and Technology, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China
Department of Radiation Oncology, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China
Department of Radiation Oncology of the Thorax Cancer (5th Radiation Oncology) Oncology, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China
Department of Radiology, Wayne State University, Detroit, United States

A R T I C L E   I N F O

Keywords:
Radiomics
MR-Linac
Adaptive radiotherapy
Robustness
Phantom study

A B S T R A C T

Purpose: Quantitative radiomics features extracted from medical images have been shown to provide value in predicting clinical outcomes. The study for robustness and reproducibility of radiomics features obtained with magnetic resonance image guided linear accelerator (MR-Linac) is insufficient. The objective of this work was to investigate the stability of radiomics features extracted from T2-weighted images of MR-Linac for five common effect factors.

Materials and method: In this work, ten jellies, five fruits/vegetables, and a dynamic phantom were used to evaluate the impact of test-retest, intraobserver, varied thicknesses, radiation, and motion. These phantoms were scanned on a 1.5 T MRI system of MR-Linac. For test-retest data, the phantoms were scanned twice with repositioning within 15 min. To assess for intraobserver comparison, the segmentation of MR images was repeated by one observer in a double-blind manner. Three slice thicknesses (1.2 mm, 2.4 mm, and 4.8 mm) were used to select robust features that were insensitive to different thicknesses. The effect of radiation on features was studied by acquiring images when the beam was on. Common movement images of patients during radiotherapy were simulated by a dynamic phantom with five motion states to study the motion effect. A total of 1409 radiomics features, including shape features, first-order features, and texture features, were extracted from the original, wavelet, square, logarithmic, exponential and gradient images. The robustness and reproducibility features were evaluated using the concordance correlation coefficient (CCC).

Result: The intraobserver group had the most robust features (936/1079, 86.7%), while the group of motion retest and radiation groups was 1072 of 1312 (81.7%) and 810 of 936 (40.0%). The stability of features in the test-retest and radiation groups was 1072 of 1312 (81.7%) and 810 of 936 (86.5%), respectively. Overall, 25 of 1409 (2.4%) radiomics features remained robust in all five tests, mostly focusing on the image type of the wavelet. The number of stable features extracted from when the beam was on was less than that extracted when the beam was off. Shape features were the most robust of all of the features in all of the groups, excluding the motion group.

Conclusion: Compared with other factors fewer features remained robust to the effect of motion. This result emphasizes the need to consider the effect of respiration motion. The study for T2-weighted images from MR-Linac under different conditions will help us to build a robust predictive model applicable for radiotherapy.

Abbreviations: MR-Linac, magnetic resonance image guided linear accelerator; CCC, concordance correlation coefficient; CT, computed tomography; PET, positron emission tomography; MRI, magnetic resonance imaging; ART, adaptive radiation therapy; ROI, region of interest; IBSI, Image biomarker standardization initiative; GLRLM, gray-level run-length matrix; GLCM, gray level cooccurrence matrix; GLSZM, gray level size zone matrix; NGTDM, neighboring gray tone difference matrix; GLDM, gray level dependence matrix; DSC, Dice similarity coefficient.

E-mail addresses: zhenjli@seu.edu.cn (Z. Li), baoshli1963@163.com (B. Li).

https://doi.org/10.1016/j.ejmp.2022.03.002
Received 31 October 2021; Received in revised form 7 February 2022; Accepted 4 March 2022
Available online 11 March 2022
Introduction

On medical images, tumors are commonly described with qualitative descriptors (microcalcification, burr, heterogeneous, cavitated) [1]. However, there has been a recent increase in quantitative, precise methods for describing the changes in a developing or radiated tumor. One-dimensional size, density, speed of growth, and so on are traditional quantitative methods for describing tumors [2]; radiomics has emerged as a newer quantitative method that allows for a more thorough extraction of large amounts of radiographic features from images [3,4]. This process has been studied as a means of providing specific imaging biomarkers and has contributed to disease classification in patients. Some studies have shown that extracted radiomics features from computed tomography (CT) [5–7], positron emission tomography (PET) [8,9], and magnetic resonance imaging (MRI) [10] can be used to predict the survival rate [11], tumor toxic reaction [12], early assessment treatment response [13–15] and lymph node metastasis [16,17].

Recently, the development of radiomics has contributed to adaptive radiation therapy (ART) [13,18].

ART is intended to provide essential insight into the real dose delivered to patients and to provide valuable information for clinical decisions. Commonly used images in ART include cone-beam CT (CBCT) and megavoltage computed tomographic (MVCT) images, which can be used for position verification during radiotherapy. In recent years, many studies have been published in the environment of CBCT and MVCT imaging, emphasizing the challenges of reproducibility of radiomic features when using different scanner acquisition parameters [19,20]. Conversely, as an emerging technology, there have been fewer discussions of the robustness of radiomics features for MRI in radiotherapy, and previous related studies have focused almost entirely on conventional diagnostic MRI [21]. Several experiments have studied the feature robustness of MR images in test–retest, interobserver and intraobserver, acquisition parameters, scanners, and so on [10,22–27]. However, compared to conventional MR, MR-Linac is constructed which integrates

![Image](https://example.com/image.png)

*Fig. 1. Workflow of the study. Here, 0° and 90° refer to the angles of the gantry beam; 0 A, 5 A, 8 A and 15 A in part ⑤ indicate the motion amplitude of the dynamic phantom. 0 vs. 5, 0 vs. 8, and 0 vs. 15 were means used to calculate the consistency between 0 mm amplitude and 5, 8, and 15 mm amplitudes respectively. ① to ⑤ refer to five influential we studied in this work. ① test-retest effect; ② intraobserver effect; ③ thickness effect; ④ radiation effect; ⑤ motion effect, including the study of motion combined with irradiation. F1, F2, F3 refer to the features selected under different conditions.*
a 1.5 T MRI closed bore MRI scanner (Philips, Best, The Netherlands) with a 7MV linear accelerator (Elekta AB, Stockholm, Sweden), mounted on a ring-shaped gantry. Recently, several academics have explored the robustness of quality for MR-Linac compared with conventional MRI in imaging systems [28,29]. However, no one has investigated the factors that commonly affect the reliability of radiomics features in MR images from MR-Linac during radiotherapy.

Thus, the main objective of this phantom study was to investigate the robustness and reproducibility of radiomic features for the frequently used MRI sequences in MR-Linac under five different influences that are commonly affected during radiotherapy and to propose robust features, which can be dependably used in future clinical studies.

Methods and materials

The workflow of the study is shown in Fig. 1.

Phantom and MR parameters

Ten jellies, five fruits/vegetables, and a dynamic motion phantom (CIRS 008Z) served as our radiomics phantoms. The different jellies and fruits/vegetables were used to reflect tumors with different shapes and texture performance and allowed for T2 weighted (T2w) images analysis and the extraction of radiomics features in the difference signal intensity images (Fig. 2). Each jelly was produced by a 0.1% water solution of agar (chembase, China). Since the jellies/fruits/vegetables are too small to imaging, a water phantom was designed in-house to assist in

Fig. 2. (a) Images of the radiomics phantom that we chose in the study. A ~ J are the ten jellies, and K ~ O are the fruits/vegetables, including one mangosteen (K), one tomato (L), one lemon (M), one onion (N), and one carrot (O). (b) The water phantom used in our work and the red box is the cavity for the jellies and fruits/vegetables. (c) An exemplarity 3-dimensional slice after segmentation with colored segmentation label. (d) The 2-dimensional segmentation label.
acquiring MR images (the volume of each jelly and fruit/vegetable is shown in Supplementary Material Table 1). The water phantom is a thorax-shaped container filled with a solution of 0.4 mM MnCl$_2$ to simulate the relaxation time T2 of muscle tissue surrounding a thoracic tumor. Five fruits/vegetables and ten jellies were placed in the cavity (Fig. 2 b) to provide MR signal and texture consistent with a set of representative thorax lesions.

The details of the dynamic phantom are shown in Fig. 3. This dynamic phantom body has a width of 32 cm, a height of 18 cm, and a length of 25.6 cm (Fig. 3 a, d). The dynamic phantom has a moving rod (length of 22 cm, diameter of 6.3 cm, Fig. 3 b) mode of liver-equivalent material that can be installed to move cyclically with varied breath-simulated curves. Due to the “tumor” in the moving rod being non-replaceable, thus the rod was replaced by a bottle (length of 20 cm, diameter of 6.3 cm, Fig. 3 c) that was the same diameter as the moving rod and fixed with the actuator using adhesive tape. To study the robustness of features influenced by different motions, we glued jellies and fruits/vegetables in the bottle one after another, moving in varying amplitudes (Fig. 3 d).

MR images were used for the phantoms in the default thorax protocol using T2 3D Tra sequence with the following parameters: TR/TE = 2100 ms/206 ms, voxel = 2.0 × 2.0 × 2.4 mm, FOV = 320 × 448 × 300 mm, scan time = 200 s-228 s, matrix = 160 × 224 × 250 slices; thickness = 2.4; and NSA = 2. The thickness parameter was manipulated in parts of the experiment to investigate whether it affected the robustness of radiomics features. In this work, radiomics features extracted from 3D images, since the ratio of the layer thickness to the pixel distance in the layer is not 1, the image data are resampled and corrected according to the actual ratio to prevent deviations in the radiomics analysis, therefore eliminating the anisotropy of the data format itself.

### The definition of CCC and selected tumor volume dependence features

The concordance correlation coefficient (CCC) is a commonly used method to evaluate the agreement between paired data [30]. In this work, the CCC was determined pairwise for features extracted from the paired images, concerning formula (1). We chose 0.9 as our cutoff value by the McBride criteria, which state that a correlation of 0.9 reflects the medium-consistency intensity, and all correlations < 0.9 are poor [31]. The features of robustness with CCC > 0.9 were then enrolled in the following experiments. The classification of CCC values is shown in Table 1.

For given features, let $\alpha$ be a vector of that feature’s values across the phantoms’ first scan. Let $\beta$ be a vector of that feature across the phantom’s second scan; $\mu_\alpha$ and $\mu_\beta$ are the features’ means. In addition, $\sigma_\alpha^2$ and $\sigma_\beta^2$ are the variances of the features. The CCC, a commonly used method to calculate the robustness of features, is thus defined as.

$$
CCC = 1 - \frac{(\alpha - \beta)^2}{\sigma_\alpha^2 + \sigma_\beta^2 + (\mu_\alpha - \mu_\beta)^2}
$$

(1)

The Spearman correlation coefficient ($\rho$) between radiomics features and the region of interest (ROI) volume was calculated, since it has been

### Table 1

<table>
<thead>
<tr>
<th>Robustness</th>
<th>The threshold values of CCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>CCC &gt; 0.9</td>
</tr>
<tr>
<td>Good</td>
<td>0.75 &lt; CCC ≤ 0.9</td>
</tr>
<tr>
<td>Medium</td>
<td>0.5 &lt; CCC ≤ 0.75</td>
</tr>
<tr>
<td>Bad</td>
<td>CCC ≤ 0.5</td>
</tr>
</tbody>
</table>

Fig. 3. (a) Picture of the dynamic phantom. (b) Picture of the moving rod. The “tumor” is stable in the center of the rod. (c) The bottle that we chose to replace the moving rod in the dynamic phantom. Ten jellies and fruits/vegetables were glued in the center of the bottle, respectively, which was stuck to the actuator using adhesive tape. (d) Picture of the dynamic phantom body.
showed that some radiomics features embed volume information [32,33]. The ROI is mainly contoured alone in the volume outside the jellies and fruits/vegetables in the acquired images, Fig. 2 (c) and (d) show the 3D surface representation and cross section of the ROI, respectively. We calculated the Spearman correlation between each feature and the test and retest image volumes individually. Any feature with $\rho > 0.85$ in both images was removed because of that feature’s strong correlation with volume [19,34]. The cutoff value of 0.85 was referenced by Faye et al [19] who considered the $\rho > 0.85$ to be highly correlated and Zou et al [34] who reported that $\rho > 0.8$ represented a strong correlation. This part aims to decrease the features that were false positives in the following analysis. We chose a high cutoff to remove features that are sensitive to different conditions rather than ROI volume.

**Phantom Test-Retest MR images**

To assess the effect of test–retest, each jelly and fruit/vegetable was scanned twice with repositioning. The interval was within 15 min, producing two images of each jelly and fruit/vegetable. After acquiring all T2w images, an experienced oncologist contoured the ROI using the commercial software AccuContour V3.0. To test the short-term robustness on MR images, the CCC was calculated pairwise for features extracted from the twice scanned images.

**Feature extraction**

A total of 1409 radiomics features were selected from the ROIs of each MR image using the commercial software AccuContour V3.0 (http://www.manteiastech.com/index_en.html), which was used for the radiomics calculation because it allows for standardized preprocessing of the medical imaging data and to extract radiomics features in various image types. This software was developed using the opensource Python package Pyradiomics version 3.0.1, which allows for the extraction of the majority of the features defined by the Image biomarker standardization initiative (IBSI) [35]. DICOM images were loaded into the AccuContour V3.0 and the ROI was contoured in the volume outside the jellies and fruits/vegetables. Defined features from original, wavelet, square, squareroot, logarithmic, exponential, and Laplacian of Gaussian-filtered images were extracted.

The main groupings of radiomics features that were selected included shape features describing the shape and geometric properties of the ROI such as the volume, maximum diameter along different orthogonal directions, maximum surface, tumor compactness, and sphericity, for a total of 14 features. First-order features were described to describe the distribution of individual voxel values without concern for spatial relationships, for a total of 18 features; second-order features provided a measure of the spatial arrangement of the voxel intensities and hence of intralesion heterogeneity and included the gray-level run-length matrix (GLRLM), gray level cooccurrence matrix (GLCM), gray level size zone matrix (GLSZM), neighboring gray tone difference matrix (NGTDM), and gray level dependence matrix (GLDM), for a total of 1377 features. Higher-order statistical features were obtained by statistical methods after applying filters or mathematical transforms to the images such as fractal analysis, Min-kowski functionals, wavelet transform, and Laplacian transforms of Gaussian-filtered images. These mathematical transforms help to extract features from areas with increasingly coarse texture patterns [36,37]. The bin-width has a strong relationship with reproducibility in MRI intraobserver sensitivity [33]. Therefore, in this work, we used a fixed bin width of 25, which has commonly been used in the literature with high reproducibility in MRI studies [32,33,38].

To increase the MRI repeatability, we normalized the image pixel values in the same range as

$$P(X) = \frac{\text{Range} \times \frac{(s) - \min_{\text{cont}}}{\max_{\text{cont}} - \min_{\text{cont}} + 1}}{A\cap B}$$

Range refers to the number of discrete values (16, 32, 64, 128), I represents the intensity of the original image, and n is the set of pixels in the ROI.

The types of features and their subclass features are shown in Supplementary Material Table 2. The description of feature extraction and definition are listed at https://pyradiomics.readthedocs.io/en/latest/features.html.

**Intraobserver effect**

Segmentation of images into ROIs such as a tumor, normal tissue, and other anatomical structures is a crucial step for subsequent informatics analysis. Manual segmentation by expert readers is often treated as ground truth. However, it suffers from high intraobserver variability [39,40]. To investigate the effect of contouring error during radiotherapy, the ROIs were contoured twice with an experienced radiation oncologist. Furthermore, to select features robust to the effect of contours, the Dice coefficient was used to calculate the spatial overlap accuracy of the two manual contours for the images in the intraobserver cohort. The value of the Dice similarity coefficient (DSC) ranges from 0 to 1. A Dice coefficient of 1 indicates exact overlap, and a value of 0 corresponds to no overlap, concerning formula (3) (Ref. [41]). This allowed coefficient us to control any reliability bias and accurately test the effect of intraobserver analysis on the extraction of radiomics features.

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

Let A be the set of robustness features on the first contour, and B be the set of robustness features on the repeat contour.

**Thickness effect**

Different thicknesses could impact fractality and lacunarity. Decreased or increased slice thickness would produce noise on the imaging [5]. To investigate whether thickness affects features acquired from MR images, we scanned ten jellies and five fruits/vegetables using slice thicknesses of 1.2 mm, 2.4 mm (default value), and 4.8 mm in the same position. Then, the correlation coefficient was calculated of 1.2 mm and 4.8 mm with 2.4 mm separately.

**Radiation effect**

The MR-Linac combines a 1.5 T MRI system with an MV Linac system that can introduce interference, especially when the radiation beam is on (irradiation) [42]. Furthermore, scattered electrons have an increase in the surface due to the interaction between radiation and phantoms. Thus, determining which features are susceptible to beam conditions was necessary. We scanned MR images before, during, and after accelerated delivery of radiation with gantry angle of 0° and 90° (Fig. 4). CVC values were calculated for each image at both 0- and 90-degree angles with the beam on. The robust features of these MR images were identified by identifying features that remained stable regardless of the beam condition. The Wilcoxon test was used to investigate whether the differences between each image under the above influencing factors were statistically significant.

**Motion effect**

The respiratory motion and heartbeats of the patients are the main reasons for uncertainty in texture values during radiotherapy. Longer MR scan times produces more serious image artifacts than CT. To analyze the effects of motion, a dynamic phantom (CRIS 008Z) was used. For this work, the dynamic part was moved in a wave of cos4(t) with amplitudes of 0, 5, 8, and 15 mm, and the details of these waves are
shown in Supplementary Fig. 1. According to the dynamic phantom’s reference, the waveform of \( \cos^4(t) \) approximates the normal respiratory function of patients. The images acquired at different amplitudes are shown in Supplementary Fig. 2 (a ~ d), which demonstrates that the degree of motion artifact increased with increased in the amplitude of motion. The influence of radiomics features during radiotherapy was complicated; thus, we acquired images with a combination of motion and irradiation (Supplementary Fig. 2 e ~ h). AccuContour V3.0 was used to manually contour ROIs on MR images obtained from the phantom, and subsequently, features were extracted from the ROIs. The CCC values were calculated for the extracted features. The robust features were repeatable in the consistency between 0 and 5 mm, 8 mm, or 15 mm.

**Results**

**Phantom Test-Retest MR images**

A total of 1409 features were extracted from the phantom images. A total of 97 features were excluded due to the absolute values of their \( \rho \) with ROI volumes being >0.85 of the test and retest images, and thus removing these features might prevent features that volume dependent. Thus, 1312 features were included in the subsequent analysis. Of the original 1312 features, 1079 features (F1) had a CCC > 0.9 in the test–retest. These included 224/270 (82.5%) features from the First-order results, 312/360 (86.7%) features from GLCM, 154/210 (73.3%) features from GLDM, 183/240 (76.2%) features from GLRLM, and 10/10 (100%) shape features in all image types including original, wavelet, square, square-root, exponential, gradient, and logarithm (Supplementary Material Table 3). The shape features showed the highest robustness in the test–retest, followed by first-order and GLCM. GLSZM features with less robust features were used to represent the consistency of texture features.

**Intraobserver effect**

Features that were robust in test–retest have enrolled in this comparison. The Dice coefficients were calculated for the contours on each jelly or fruit/vegetable. The mean ± standard deviation Dice similarity coefficient was 0.96 ± 0.03 between the twice contour.

The robustness was consistently high for intraobserver analysis with a median value of 0.982 (range for overall pooled analysis 0.319–1.000). Overall, radiomics features showed excellent intraobserver robustness (\( n = 936/1079, 86.7\% \), F3). Shape, and first-order were the most robust features (\( n = 10/10, 100\% \); and \( n = 208/224, 92.9\% \)), in which all of the subclasses remained stable, and the NGTDM feature (\( n = 29/59, 67.4\% \)) was the least robust feature under the intraobserver effect.

**Thickness effect**

The number of robustness features was significantly different when extracted from 1.2 vs. 2.4 mm and 2.4 vs. 4.8 mm slice images. The features extracted from 1.2 vs. 2.4 mm slice thickness were better than from 2.4 vs. 4.8 mm for all feature types. Features that were robust in test–retest and intraobserver analyses were enrolled this comparison. The robustness of the feature was lower for 2.4 vs. 4.8 mm thicknesses (\( n = 416/936, 44.4\% \)) than for 1.2 vs. 2.4 mm thicknesses (\( n = 591/936, 63.1\% \)). According to all feature classes, shape features were most robust across all scanning thicknesses (\( n = 10/10, 100\% \)). GLCM features type extracted from 1.2 vs. 2.4 mm thickness showed higher sensitive robustness as compared with those extracted from the 2.4 vs. 4.8 mm thickness images (1.2 vs. 2.4 mm thickness \( n = 168/360, 46.7\% \); 2.4 vs. 4.8 mm thickness \( n = 107/360, 29.7\% \)). A total of 374/936 (40.0%) features showed excellent robustness across the different thicknesses. The value of the CCC in 1.2 vs. 2.4 mm thickness was mainly >0.9, while the value of the CCC was concentrated in the group with a range of 0.9 to 0.75 in 2.4 vs. 4.8 mm thickness (Fig. 3 (a) in supplement).

**Radiation effect**

Across all feature types, the number of stable features was mostly higher for features calculated from the beam on images than for those calculated from the beam off images (Fig. 3 (b) in Supplement). The images acquired from when the beam was on provided a higher percentage of robust features in different gantry angles. Features that were robust in test–retest and intraobserver analyses passed this comparison. For a 0° gantry angle, the robust features calculated from the images when the beam was on were 883/936 (94.3%) and 896/936 (95.4%)
features when the beam was off. With the gantry angle positioned at 90°, 876/936 (93.6%) features exhibited robust when the beam was on, and 878/936 (93.8%) features kept robustness when the beam was off. In addition to the effect of radiation, the robustness of features can change with different beam positions, with 793/936 (84.7%) features remaining robust when the beam position was changed from 0° to 90° when the beam was on. The details are shown in Supplementary Material Table 3. The Shape and NGTDM features were both the most stable features with the beam both on and off, regardless of the beam position. The intersecting set of features in different conditions is shown in Fig. 3 (c) in the Supplement. A total of 810/936 (86.5%) features maintained robustness with the beam both on and off, with the least number of stable features seen in the GLSZM group. The greatest number of subclass features in NGTDM were complexity features.

After the consistency correlation analysis, the Wilcoxon test was used to perform statistical analysis on the features obtained under different beam conditions (Supplementary Material Table 4). The P-value between beam on or off were mostly less than 0.0001.

Motion effect

In this part, we examined whether changing the amplitude of motion produced changes in the robustness of radiomic features. Features that were robust in test–retest and intraobserver tests were passed this comparison. Among all of the features, the number of radiomics features robustly decreased with increasing motion amplitude, (Fig. 5(a)). Three features, DependenceNonUniformity, GrayLevelNonUniformity, and RunLengthNonUniformity were robust for a motion of 5–15 mm. For motion with 5 mm amplitude, n = 158/936 (16.9%) features were stable. At 8 mm of motion, n = 112/936 (12.0%) features were robust, and n = 73/936 (7.8%) features were repeated at an amplitude of 15 mm. After combining radiation with motion, 131/936 (13.9%), 90/936 (9.6%), and 30/936 (3.2%) features were remained robust in 5, 8, and 15 mm amplitudes, respectively, and the number of robustness features of the combined effect of radiation and motion was less than those influenced by motion, Fig. 5 (a), with details as shown in Table 3 in the Supplement. Under this condition, four features, GrayLevelNonUniformity, RunLengthNonUniformity, Energy and TotalEnergy, remained stable for amplitudes of 5–15 mm.

Compared with other perturbations, motion most affected the

![Fig. 5.](image-url)

Fig. 5. (a) Results of the motion test and combined motion with irradiation. M refers to the effect of motion; RM refers to the effect combining motion with irradiation. Larger values refer to better consistency with 0 mm amplitude in which the phantom was stable. Heatmap illustrating the changes inconsistency of the changes with the increased amplitudes with or without irradiation. (b) The CCC varieties between pairs of amplitudes were plotted by the types of groups being compared. Here, 0 vs. 5, 0 vs. 8, and 0 vs. 15 were means used to calculate the consistency between 0 mm amplitude and 5, 8, and 15 mm amplitudes.
stability of radiomics features and emphasized the importance of reducing the motion effect during radiotherapy to enhance image robustness. Fifty-six features were reproducible for the effect of motion. Fourteen features remained robust to both the effect of motion and the influence of combined motion with radiation, focusing on the image type of the wavelet. GrayLevelNonUniformity was the most robust feature among these features, and was used to measure the changes in gray intensity and run length. The results are shown in Fig. 5(b), where the concordance correlation coefficient for GrayLevelNonUniformity from GLDM and GLRLM also demonstrates the illustrative conclusion. Only the coarseness features, which were used to represent the unevenness of images, remained robust in the group of NGTDM features.

Multivariate robust features are defined as features that have excellent robustness under different types of outside influences. Under the effects of five factors, the feature changes are shown in Fig. 6 (a). The intraobserver features remained most robust in the whole test (936/1079, 86.7%), followed by the effect of radiation (810/936, 86.5%). The number of robustness features under the influence of motion was less than that of other factors. The changing trend in the GLSZM feature was similar to the various of GLDM. The GLCM feature type described by the spatial correlation was the most sensitive of these factors. The intersecting set of robust features of all the effect factors is shown in Fig. 6 (b). A total of 1.9% (25/1312) of features remained robust in the whole test, mainly focusing on the original image type and wavelet image type. The details are shown in Table 5 in the supplement. Motion was the most influential factor on the robustness of radiomics features (Fig. 6 c).

Discussion

As new and advanced radiotherapy equipment, MR-Linac makes daily MRI scanning possible for treatment response monitoring [43]. However, the magnetic field intensity (1.5 T) and the linear accelerator render it different from the conventional diagnostic magnetic resonance system [44]. Therefore, the robust radiomics features selected from conventional diagnostic MRI are not suitable for MR-Linac. In this work, we systematically investigated the effects of test-retest, intraobserver, thickness, radiation, and motion robustness on radiomics features derived from T2w images commonly used in conventional clinical practice in a well-controlled phantom setup, which could be helpful to establishing a reliable clinical prediction model.
The test–retest cohort provides a controlled environment to determine the radiomics features most likely to determine the inherent characteristics of the tumor [45]. Like in other studies, we observed a significant effect of test–retest variability on the extracted radiomics features. Although we used the same scanning parameters and with position repeats, many radiomic features appear to be highly sensitive to machine influence. Therefore, removing the features sensitive to test–retest is necessary for the following radiomics investigation, laying a foundation toward the establishment of clinical modes. In this study, 82.2% of robust features in the test–retest remained robust, which was better than other studies. For example, Baeßler et al. [10] showed that 46.0% and 54.0% of features maintained robustness on low-resolution and high-resolution T2w images, respectively. Shiri et al. [46] demonstrated that 74.0% of assessed radiomics features on MRI, had high test–retest stability in thirteen glioblastoma patients.

The intraobserver as the most essential step of the workflow of radiomics, more easily led to bias in tumor delineation from delineation subjectivity of observers. In this work, 86.7% of features maintained excellent robustness in the intraobserver cohort, indicating that the highest robustness in all factors was similar to the result of Dreher et al. [27] who indicated that the intraobserver and interobserver were high for all sequences that they studied. Similarly, a previous MRI phantom study reported that robust intraobserver features were superior to other factors [10]. Reducing the sensitive features influenced by subjective bias through automatic image segmentation is a major step for radiomics research, leading toward the development of automated image segmentation and artificial intelligence [47].

Previous studies have investigated the effects of thickness on tumor size measurements (e.g., volume) in cancer screening programs and therapy response assessment [48]. To our knowledge, our study was the first to investigate the effects of MR slice thickness from MR-Linac on quantitative image radiomics features extracted to denote the tumor’s volume, shape, and density using a manageable phantom. The different thickness cohorts had the largest effect of radiomics features in our work, with only 40.0% of the features remaining stable. Our results showed that with increasing thickness, fewer features retain stability. This result is similar to Zhao et al. [5] who showed that a thinner slice thickness was better than a thicker slice thickness for density mean, density 3D gray-level cooccurrence matrix (GLCM) energy, and homogeneity. In our consideration, a larger thickness caused a sufficient number of voxels to not be reached, which could explain the low robustness found for this factor. The thickness of daily images is constant and easy to control; therefore, maintaining the consistency of image thickness remains an indispensable measure to ensure the stability of features during radiotherapy.

Despite considerable changes in the number of robust radiomic features between various beam conditions, we showed that 810 features were exceedingly robust to variations between beams being on and off, and between gantry angle changes. The system of the accelerator was the main difference between conventional MR and MR-Linac, consequently, the effect of beam conditions should be considered. Electrons from radiation cause a larger amount of noise on the images, as well as artifacts that influence the quality of such images [49]. As a result, the number of reproducibility features extracted from the beam on was fewer than those extracted from the beam off with the gantry at the 0° and 90° positions. Our results, in this case, differ from those of Wang et al. [28], who demonstrated that there were no significant differences in image robustness or quality with or without radiation.

Motion, as the leading bias originating from respiration and heartbeat during radiotherapy, the robustness and variability should be the first consideration when investigating the robustness of radiomics features extracted from tumors in the thorax and abdomen [49]. For MRI, the imaging time was longer than that for CBCT. Thus, for thoracic and abdominal tumors, the effect of respiratory and heartbeat motion on the robustness of radiomics features extracted from MRI was lower than that from CBCT images during radiotherapy. Fave et al. [19] reported that patients with NSCLC had tumor motion ranging from 5 mm to 10 mm, and they also claimed that most features changed substantially with the increasing amplitude of the ROI. Nevertheless, in this study, to identify features that were less affected by movement, a larger range of motion was chosen. The larger motion amplitude might filter out some features that could be used in the future but should determine more robust features for clinical applications in chest or abdomen tumors.

From this study, we can come to four key points. First, shape features are the most stable and reproducible features in all tests except the effect of motion. Similarly, this point has been reported by several works in the literature, not only from patient data but also from phantom studies [10, 45, 50]. Second, considerable differences exist in the number of robust features between different beam conditions. The number of robustness features extracted from the beam off images was greater than that from the beam on. Third, the effect of intraobserver draws the highest percentage of robust features in this phantom study, followed by the beam condition. Fourth, 25 of 1409 features remained robust to the variations among all five tests in our study (Supplement Table S). Therefore, we suppose that these features can be used to design radiomics features within clinical studies.

Our study has several limitations. First, although we used jellies and fruits/vegetables of various shapes to mimic patient tumors, the irregular jellies and fruits/vegetables were fundamentally different from real tumors, and the edge of the jellies and fruits/vegetables were smooth, which does not mimic common tumor features such as microcalcifications, burrs and lobular configurations, which are all common in malignant lesions that would require radiotherapy in the first place. This tumor-alternative method might show variability in phantom textures in five common factors during radiotherapy. Due to the different textures inherent to the actual tumors, the features that we extracted cannot fully reflect the true tumor situation. Second, the radiomics features were extracted from a phantom study. These phantoms were different from the tissue composition of patients’ bodies, and this was a single-center phantom study that might not be investigated by other machines. Furthermore, we investigated the most commonly used MRI sequence and did not consider other quantitative MRI sequences, such as T1-weighted imaging and fluid-attenuated inversion recovery (FLAIR) imaging. In the future, we wish to verify the robustness features from some patients in different sequences.

In conclusion, the robustness and reproducibility features of MR-Linac were tested in this work under five common influencing factors, and thus can be reliably applied in clinical studies. Shape features are considered the most robust features except for the influence of motion. The number of robustness features of the motion cohort was the fewest, emphasizing the need to decrease the amplitude of respiration motion. The 25 features with excellent robustness under five influences could be most acceptable for the design of future radiomics signatures, although these results should be validated in the clinical study in cancer patients.

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.

Acknowledgements

This work was supported by Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical. Support was also provided by the National Key Research and Development Program of China (2016YFC0105106); Academic Promotion Program of Shandong First Medical University(2019LJ004); The National Natural Science Foundation of China (82102173). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies.
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.epmmed.2022.03.002.

References


