Radiotherapy and Oncology 135 (2019) 107–114

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Orignal Article

Repeatability and reproducibility of MRI-based radiomic features in cervical cancer

Sandra Fiset [a], Mattea L. Welch [f,a], Jessica Weiss [e], Melania Pintilie [e], Jessica L. Conway [a], Michael Milosevic [a,b], Anthony Fyles [a,b], Alberto Traverso [a,h], David Jaffray [a,b,g], Ur Metser [c,d], Jason Xie [a], Kathy Han [a,b,*]

[a] Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto; [b] Department of Radiation Oncology, University of Toronto; [c] Joint Department of Medical Imaging, University Health Network, Toronto; [d] Department of Medical Imaging, University of Toronto; [e] Department of Biostatistics, University Health Network, Toronto; [f] Department of Medicine, Medical Biophysics, University of Toronto; [g] Quantitative Imaging for Personalized Cancer Medicine, Techna Institute, University Health Network, Canada and [h] Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, the Netherlands

ARTICLE INFO

Article history:
Received 21 November 2018
Received in revised form 2 March 2019
Accepted 4 March 2019

Keywords:
Radiomics
MRI
T2-Weighted
Cervical cancer
Repeatability
Reproducibility

ABSTRACT

Purpose: The aims of this study are to evaluate the stability of radiomic features from T2-weighted MRI of cervical cancer in three ways: (1) repeatability via test–retest; (2) reproducibility between diagnostic MRI and simulation MRI; (3) reproducibility in inter-observer setting.

Materials and methods: This retrospective cohort study included FIGO stage IB-IVA cervical cancer patients treated with chemoradiation between 2005 and 2014. There were three cohorts of women corresponding to each aim of the study: (1) 8 women who underwent test–retest MRI; (2) 20 women who underwent MRI on different scanners (diagnostic and simulation MRI); (3) 34 women whose diagnostic MRIs were contoured by three observers. Radiomic features based on first-order statistics, shape features and texture features were extracted from the original, Laplacian of Gaussian (LoG)-filtered and wavelet-filtered images, for a total of 1761 features. Stability of radiomic features was assessed using intraclass correlation coefficient (ICC).

Results: The inter-observer cohort had the most reproducible features (95.2% with ICC $\geq$0.75) whereas the diagnostic–simulation cohort had the fewest (14.1% with ICC $\geq$0.75). Overall, 229 features had ICC $\geq$0.75 in all three tests. Shape features emerged as the most stable features in all cohorts.

Conclusion: The diagnostic–simulation test resulted in the fewest reproducible features. Further research in MRI-based radiomics is required to validate the use of reproducible features in prognostic models.

 Radiotherapy and Oncology 135 (2019) 107–114

Radiomics, the automated high-throughput extraction of quantitative imaging features, is hypothesized to capture the histological heterogeneity inherent to solid tumors [1–3]. The potential of radiomics has instigated a multitude of modality- and site-specific investigations to provide robust diagnostic and prognostic models. Generally, computed tomography (CT)-based radiomics have dominated the literature; however, magnetic resonance imaging (MRI) is gaining popularity owing to its superior soft tissue contrast [4].

While radiomics is changing the landscape of cancer imaging research, the lack of consistency in analysis and feature reporting make comparison and repetition of studies difficult [5]. Consequently, studies looking at the repeatability (comparison under constant condition) and reproducibility (comparison under varying conditions) of radiomic features have become increasingly common [6]. Identification of reproducible and repeatable features, and their inclusion in predictive models, are key to ensuring model generalizability.

An important indicator of feature repeatability is test–retest, a comparison of radiomic features from two images of the same patient acquired within a short timeframe. Studies looking at two sets of CT images acquired within 15 minutes to 2 weeks found that 29%–98% of calculated features were not repeatable, thus confirming the need for robust feature selection [7–14]. There have been no conclusive studies regarding the test–retest robustness of MR-based radiomic features.

Abbreviations: ICC, intraclass correlation coefficient; LoG, Laplacian of Gaussian; GLCM, gray-level co-occurrence matrix; GLSZM, gray-level size zone matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run length matrix; NGTDM, neighboring gray tone difference matrix; NSCLC, non-small cell lung cancer.

* Corresponding author at: Department of Radiation Oncology, Princess Margaret Cancer Centre, University Health Network, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada.

E-mail address: Kathy.Han@rmp.uhn.ca (K. Han).

https://doi.org/10.1016/j.radonc.2019.03.001
0167-8140/

In addition to a diagnostic MRI, patients planned for radiotherapy often undergo a simulation MRI in treatment position for radiation treatment planning using a different MRI scanner and image acquisition protocol. Clinical applicability of radiomics will be dependent on its widespread external generalizability. It is therefore essential to identify radiomic features that are able to transcend such differences between image acquisition parameters.

Additionally, tumor delineation uncertainty can translate into significant variability in radiomic feature accuracy [15,16]. The need for assessment of inter-observer variability in MRI radiomics is further substantiated by two published studies which have shown better reproducibility than CT [17,18].

Ultimately, there is a need to identify MRI-based radiomic features that are robust and stable against inevitable variation in clinical data. We hypothesize that we will identify MRI-based radiomic features that are robust to tests of repeatability and reproducibility, which can be utilized in predictive radiomics models. Accordingly, the aims of this study are to evaluate the stability of radiomic features from T2-weighted MRI of cervical cancer in three ways: (1) repeatability via test–retest; (2) reproducibility between diagnostic MRI and simulation MRI; (3) reproducibility in inter-observer setting.

Materials and methods

Study population

This retrospective cohort study was approved by the institutional research board, with waiver of informed consent. We retrospectively identified all patients with stage IB-IVA cervical cancer who were treated at our center with chemoradiation between 2005 and 2014. Those who did not undergo diagnostic MRI at our center prior to treatment were excluded. There were three cohorts of women: (1) 8 women who underwent test–retest simulation MRIs (within 14–47 min); (2) 20 women who underwent a diagnostic MRI and a simulation MRI within an average timeframe of 8 days; (3) 34 patients whose diagnostic MRIs were contoured by three observers (Fig. 1). There was overlap between the three patient cohorts. Table 1 outlines patient demographics.

Image acquisition

All images were acquired on clinical MR scanners with axial T2-weighted turbo spin-echo (TSE) sequence. Scanner and imaging parameters are listed in Table 1. All imaging parameters were the same between images from a single patient in the test–retest cohort and the inter-observer cohort. The diagnostic–simulation cohort had differences in imaging parameters for a given patient including scanner model, magnetic field, echo time (TE) and repetition time (TR), as is expected in real clinical scenarios.

The cervical tumor was manually delineated on all images by one gynecologic radiation oncologist (KH) with 5 years of experience using Raystation 6 (RaySearch Laboratories). To minimize intra-observer contouring variability between diagnostic and simulation MRIs in the test–retest cohort, each patient's images were soft-tissue co-registered. Co-registration involved contouring the diagnostic T2 MRI, propagating the contour onto the simulation MRI, and modifying as needed. Additionally, the inter-observer cohort was subsequently contoured by two other gynecologic radiation oncology observers (JC and JX with 1 and 10 years of experience, respectively).

Feature extraction

After contouring, all DICOM images and associated contours were exported and resampled to $0.6 \times 0.6 \times 4$ mm to exclude

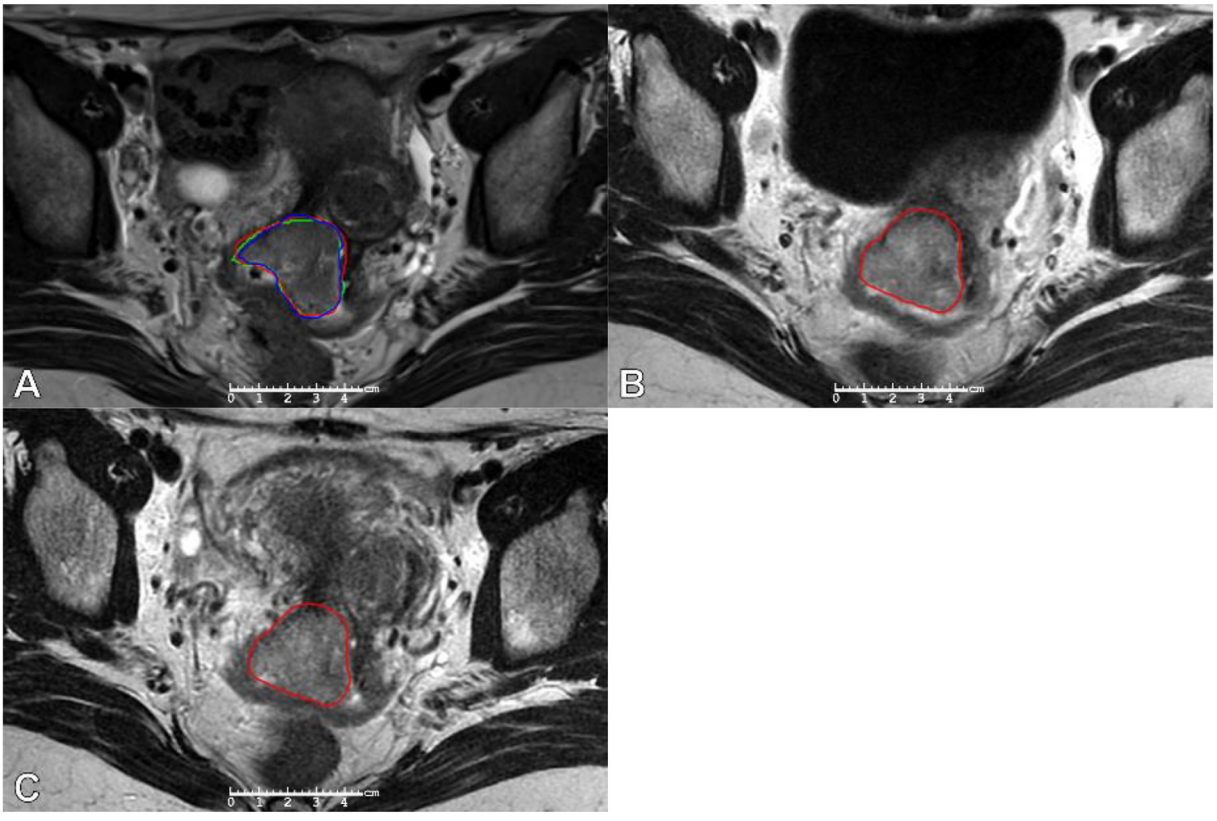

Fig. 1. Axial T2-weighted MR images of a patient with FIGO stage 2B cervical cancer. (A) Diagnostic MR images with contours by three observers, (B) Radiotherapy simulation MR images acquired on a different scanner approximately 2 hours following A and (C) Radiotherapy simulation MR images acquired 20 minutes following B on the same scanner, after a bathroom break.

Table 1
Patient demographics and image parameters for the three patient cohorts in this study (test–retest, diagnostic–simulation, inter-observer).

Characteristics	Patient cohorts Test–retest	Diagnostic–simulation		Inter-observer
Number of patients	8	20		34
Age, mean ± SD *(yrs)*	54 (9.4)	51 (8.8)		49 (26 – 70)
Tumor characteristics				
FIGO Stage *(n)*				
1B	4	8		13
2A	2	2		5
2B	1	4		10
3A	0	1		0
3B	1	5		6
Max dimension on imaging, mean ± SD *(cm)*	4.1 ± 1.4	4.8 ± 1.6		5.0 ± 1.5
Volume, mean ± SD *(cm^3)*	29.2 ± 21.0	49.5 ± 44.1		46.0 ± 33.3
Time between images, median (range)	23 min (14–47 min)	8 days (0–14 days)		N/A
Image Parameters	Test–Retest*	Diagnostic	Simulation	Inter-observer
Number of images *(n)*	8	20	20	34
MR Scanner *(n)*				
GE Signa Excite	0	12	2	20
GE Signa HDx	7	1	14	0
Siemens Verio	1	3	2	9
Siemens Avanto	0	4	2	5
Magnetic Field *(n)*				
1.5 T	7	17	18	25
3.0 T	1	3	2	9
Sequence, median (range)				
Slice Thickness *(mm)*	4	4	4	4
Axial resolution *(mm)*	0.43 (0.43–0.69)	0.43 (0.43–0.78)	0.45 (0.39–0.78)	0.43 (0.43–0.69)
TE *(ms)*	102 (92–102)	98 (88–104)	104 (92–106)	98 (88–106)
TR *(ms)*	6100 (3850–6500)	3820 (3050–6340)	5842 (3500–6500)	3790 (3100–7667)

SD = standard deviation.
* All imaging parameters were the same between images from a single patient in the test–retest cohort.

potential confounding by variable in-plane resolutions. Resampling was performed using B-spline interpolation which has been shown to retain tissue contrast differences and has good reproducibility [19,20]. Resampling and subsequent feature extraction were performed using the open-source PyRadiomics (v.1.3.0) package for Python (v. 3.6.5) [21]. The custom script which was used to run PyRadiomics is included in this paper as Supplementary materials. The PyRadiomics platform was selected for radiomic feature extraction to increase accountability and refinement of methodologies. Additionally, this platform was validated against the Image Biomarker Standardization Initiative benchmark values [22].

MRI gray values (signal intensity) are generally relative and cannot be compared between images. To ensure better comparability of gray values, normalization was performed on the images by centering at the mean and dividing by standard deviation of the gray values in the image as per PyRadiomics standard. In both the literature and the PyRadiomics documentation, a fixed bin width is recommended as opposed to a fixed bin count [23]. An analysis was performed to determine a suitable bin width value. Due to the normalization, smaller bin widths rather than the default value of 25 in PyRadiomics were required to achieve a sufficient number of bin counts for each patient. A fixed bin width of 0.05 was deemed suitable as it resulted in an average of 54 bins (minimum 17, maximum 95) in the original images. While the sources mentioned above recommend the fixed bin width method, the Image Biomarker Standardization Initiative recommends the use of fixed bin count for T2-weighted MR [22]. To further evaluate the differences between the two methods, the analysis was repeated using a fixed bin count of 64, which has been commonly used in the literature with good reproducibility in PET studies [24–26].

A total of 1761 features were computed for each image. The main groupings of texture analysis features were (1) First-order statistics based on pixel gray-level histograms, 18 features; (2) Shape metrics, 13 features; (3) Statistical features derived from texture matrices including gray-level co-occurrence matrix (GLCM), gray-level size zone matrix (GLSZM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), neighboring gray tone difference matrix (NGTDM), 74 features; (4) Statistical features derived from texture matrices in Laplacian-of-Gaussian (LoG) filtered domain (0.5–5.0 mm kernels), 920 features; and (5) Statistical features derived from texture matrices in wavelet filtered domains, 736 features. Texture matrices were calculated in 3 dimensions, resulting in 2 neighbors for each of 13 angles. As per PyRadiomics default, feature values are calculated in all directions and the mean was recorded. No weighting to distance was applied to the GLCM matrix.

Statistics

Feature stability was evaluated using the Intraclass correlation coefficient (ICC). ICC(1,1) was used for the test–retest and diagnostic–simulation cohorts, whereas ICC(2,1) was used for the inter-observer cohort. Here, an ICC of $\geq$0.75–0.89 was considered good reproducibility and an ICC $\geq$ 0.90 was considered excellent reproducibility as recommended by Koo et al. [27]. The Dice coefficient was used as a metric for the spatial overlap accuracy of the three manual contours for the images in the inter-observer cohort. A Dice coefficient of 0 indicated no overlap and a value of 1 corresponded to exact overlap [28,29].

Features which were highly correlated were grouped together in clusters to avoid skewing results if many features show high

ICC but, in fact, are all highly correlated and would not add additional value to a radiomics model. Cluster sizes of 10, 100, 200, 300, 400, 500, and 600 were examined. The optimal cluster size was decided to be the one which 75% of pairs of features in a cluster are correlated with a Pearson correlation coefficient above 0.9. This ensures that clusters are highly correlated within themselves but still reduces the number of features. The representative feature from each cluster was selected as the feature with the highest median correlation with the other members of the cluster.

The Pearson correlation coefficient was used to evaluate the relationship between features and tumor volume. Volume is a known prognostic indicator; therefore, features which are highly correlated with volume do not add meaningful information to a radiomics model and volume dependency can artificially increase a feature's repeatability [30].

In order to determine whether a specific LoG filter or wavelet decomposition offered superior feature stability, the first-order and texture features calculated on the original image versus the same features calculated in 19 image domains were compared. Therefore, the original image was compared with 10 images from LoG kernel sizes ranging between 0 and 5 mm, and 8 images from the wavelet decompositions. For this analysis, the difference between the ICC of the original image and each filtered image was calculated. Only the features which exhibited an ICC $\geq$ 0.5 in one of the image domains were included in the analysis to reduce artificially high differences between very low ICCs which are not of interest for potential inclusion in radiomics models.

An alternative measure of agreement, Krippendorff's alpha, was calculated to assess for reliability in the three cohorts. Krippendorff's alpha ranges from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement [31]. All statistics were performed with R package v 3.4.2, 2017.

Results

The Dice coefficients were calculated for the contours on each patient. The mean ± standard deviation Dice coefficients were 0.92 ± 0.03, 0.90 ± 0.06 and 0.91 ± 0.06 between observers 1 and 2, 1 and 3, and 2 and 3, respectively.

ICC values for the fixed bin width and fixed bin count methods for the original image domain are provided in Table E1 (online). The fixed bin width method produced higher ICCs for the inter-observer cohort whereas the fixed bin count method resulted in higher ICCs for the test–retest cohort. The two methods were approximately equal for the diagnostic–simulation cohort. Therefore, neither method emerged as superior for this study. Only results from the fixed bin width method are reported for the remainder of this text.

The number of features that fell within either the "good" ($\geq$0.75–0.89) or "excellent" ($\geq$0.9) ICC category for each cohort is presented in Table 2. The shape metrics have the highest percentage of features in the "excellent" ICC group in all three cohorts. Overall, the diagnostic–simulation cohort showed the fewest features with "good" or "excellent" ICC, 14.1% of all features. This contrasts with the test–retest and inter-observer cohorts from which 52.1% and 95.2% of features had "good" or "excellent" reproducibility.

In addition to analyzing all the features separately, features which were highly correlated were clustered. The optimal number of clusters was 300 where 75% of pairs within each cluster have a Pearson correlation of above 0.90 or below −0.90. When analyzing only 1 representative feature from each cluster, the percentage of features which demonstrated "excellent" or "good" reproducibility in the three cohorts remained largely unchanged as listed in Table E2 (online). This confirms that there is no skewing of results from highly correlated features. The diagnostic–simulation cohort again demonstrated the fewest reproducible features with 15.0% (45/300) having ICC $\geq$ 0.75. The test–retest and inter-observers cohorts showed 51.7% and 95.6% of representative features have ICC $\geq$ 0.75.

Across all three cohorts, 229 common features out of the total 1761 features (including all image domains) had a "good" ICC value $\geq$0.75, and 99 features had an "excellent" ICC value $\geq$0.9 as illustrated in Fig. 2. Of the 229 features which had ICC $\geq$ 0.75 in all three cohorts, 150 features also had a Pearson correlation coeffi-

Table 2
Number of features (n) and percentage of their groups (%) which fall into excellent ICC category (ICC$\geq$0.9), good category (ICC$\geq$0.75–0.89) and other (ICC < 0.75) for all features and distinct feature types (first-order, shape, texture, LoG filtered and wavelet filtered).

	Test–Retest		Diagnostic–Simulation		Inter-observer	
Feature Category (n)	n	%	n	%	n	%
All features (1761)						
ICC $\geq$ 0.9	398	22.6	109	6.2	1310	74.4
ICC $\geq$ 0.75–0.89	519	29.5	139	7.9	366	20.8
ICC < 0.75	844	47.9	1513	85.9	85	4.8
First-order (18)						
ICC $\geq$ 0.9	6	33.3	2	11.1	12	66.7
ICC $\geq$ 0.75–0.89	7	38.9	1	5.6	5	27.8
ICC < 0.75	5	27.8	15	83.3	1	5.6
Shape Metric (13)						
ICC $\geq$ 0.9	12	92.3	12	92.3	13	100.0
ICC $\geq$ 0.75–0.89	1	7.7	0	0.0	0	0.0
ICC < 0.75	0	0.0	1	7.7	0	0.0
Texture (74)						
ICC $\geq$ 0.9	19	25.7	4	5.4	47	63.5
ICC $\geq$ 0.75–0.89	24	32.4	2	2.7	25	33.8
ICC < 0.75	31	41.9	68	91.9	2	2.7
LoG (920)						
ICC $\geq$ 0.9	226	24.6	62	6.7	648	70.4
ICC $\geq$ 0.75–0.89	307	33.4	113	12.3	225	24.5
ICC < 0.75	387	42.1	745	81.0	47	5.1
Wavelet (736)						
ICC $\geq$ 0.9	135	18.3	29	3.9	590	80.2
ICC $\geq$ 0.75–0.89	180	24.5	23	3.1	111	15.1
ICC < 0.75	421	57.2	684	92.9	35	4.8

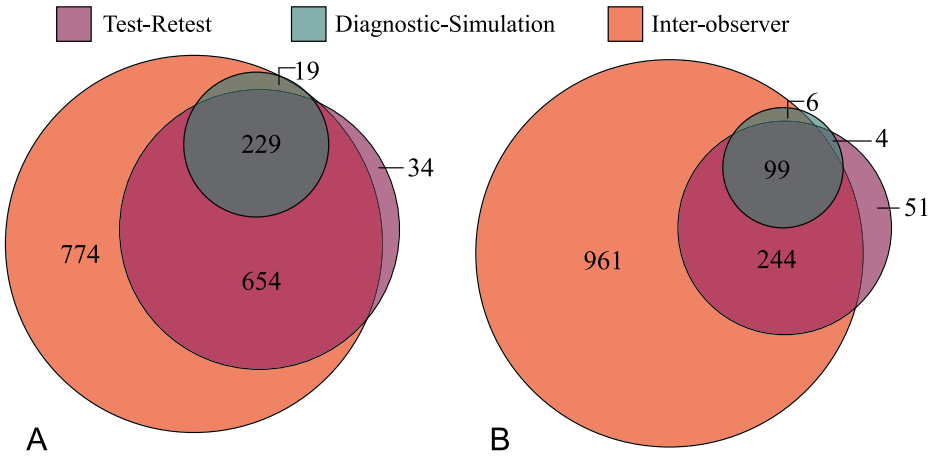

Fig. 2. (A) Venn diagram illustrating the number of features which have ICC ≥ 0.75 in the three cohorts (B) Venn diagram illustrating the number of features which have ICC ≥ 0.9 in the three cohorts.

cient of less than 0.9 with volume (i.e. not highly correlated with volume). Many of the features with both ICC ≥ 0.75 and Pearson correlation coefficient <0.9 were repeated in multiple image domains. A list of the ICC values and 95% confidence interval for all radiomic features computed is provided in Table E3 (online). Table E4 provides their Krippendorff's alpha values and Pearson correlation coefficients.

First-order and texture features were calculated in 19 image domains: the original image, 10 images with LoG kernel sizes ranging between 0 and 5 mm, and 8 images from the wavelet decompositions. To explore any variation in feature stability (ICC) by *image domains*, ICCs for the 13 first-order and 74 texture features were combined by image domain in Fig. 3. The ICCs from the features for the original image are included in each graph for comparison. The diagnostic–simulation cohort demonstrates significantly lower ICCs in all image domains. The range of ICC values varies between image domains with no clear image domain emerging as superior to the others.

To compare feature stability in the *original versus filtered images*, the differences between the ICCs for each feature calculated on the original image and each filtered image were plotted for each of the three patient cohorts (Fig. 4). The LoG filtered images showed better ICCs than the original image for the diagnostic–simulation cohort, and worse ICCs for the test–retest and inter-observer cohorts. For the diagnostic–simulation cohorts, 31.9% of features demonstrated >10% higher ICCs with LoG filtered images when compared to the original image in contrast to 23.1% which showed ≥10% lower ICCs with LoG filtered images. The test–retest and inter-observer cohorts on the other hand demonstrated 24.4% and 3.4% of features with ICCs >10% higher in LoG filtered images, and 28.3% and 5.1% of features which demonstrated ICCs ≥ 10% lower with LoG filtered images, respectively. The original image demonstrated better ICCs than the wavelet filtered images in the diagnostic–simulation and test–retest cohorts. Respectively, 70.1% and 39.0% of features had ≥10% higher ICC in their original images when compared with wavelet filtered images. This is compared to 6.9% and 22.4% of features which had >10% lower ICC in their original image when compared with wavelet filtered images in the same cohorts. The inter-observer cohort demonstrated modestly higher ICCs from wavelet filtered images than from the original image domain (10.8% vs 5.6%). Further breakdowns of feature differences between original images and filtered images categorized by filter or texture feature type are supplied in Fig. E1 A-L.

Discussion

Radiomics has emerged as a means of image-based prognostication. Ensuring radiomic feature stability is imperative to the external generalizability of such prognostic models. It is anticipated that this study will help guide the selection of stable radiomic features

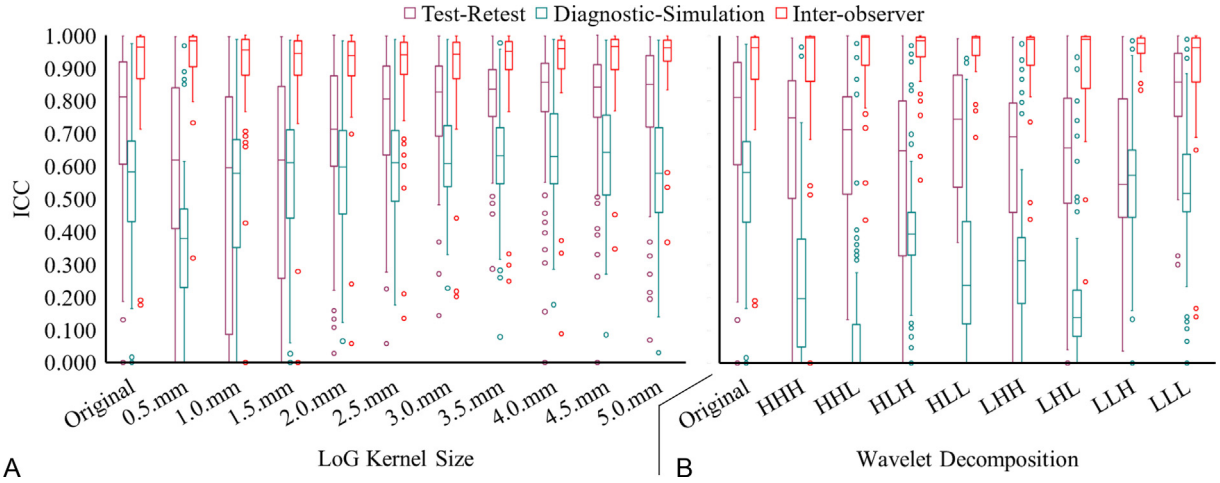

Fig. 3. Box plot illustrating the distribution of intraclass correlation coefficients (ICC) for the first-order (n = 13) and texture features (n = 74) derived from the original image, Laplacian of Gaussian (LoG) filtered images with kernel sizes 0.5–5.0 mm and each wavelet decomposition.

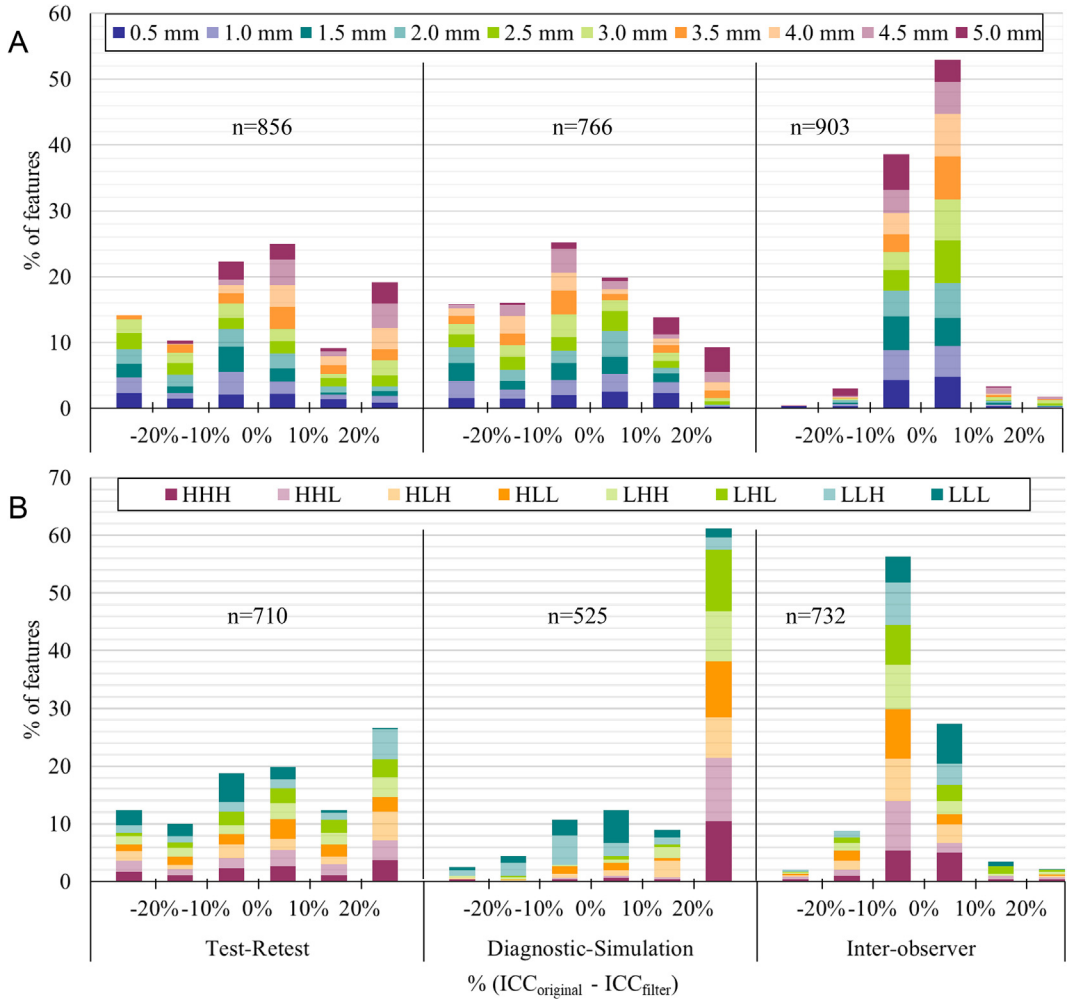

Fig. 4. Histogram demonstrating the percentage difference between the ICC for the original vs. filtered feature for the three study cohorts. Only features which have ICC > 0.5 in either the original image or the filtered image ("n" on the figure) are included. Each color in the bars represents the ICC differences between the original image and individual filters. (A) The ICC differences between the original image and the various LoG kernel size filters. (B) The ICC differences between the original image and the wavelet decompositions.

in future prognostic models by evaluating feature repeatability and reproducibility of radiomic features in three tests. Specifically, the test–retest cohort offers a controlled environment to identify radiomic features which most likely identify characteristics inherent to the tumor. The diagnostic–simulation cohort aims to identify features which are robust against differences in scanners and acquisition protocols, thus mimicking a clinical scenario. Both the test–retest and diagnostic–simulation evaluate errors originating from data acquisition. The inter-observer cohort, on the other hand, evaluates error originating from tumor delineation, another important clinical scenario. Combining the results from the three cohorts can represent a good strategy to perform feature dimensionality reduction.

The importance of careful feature selection is first demonstrated in the test–retest cohort which resulted in 47.9% of the features with ICC <0.75 (below good reproducibility) despite the controlled setting. Likewise, even with the use of a phantom and identical imaging parameters, one study has shown that 4% of CT-radiomic features had a concordance correlation coefficient (a numerically similar but alternative popular agreement index to ICC which does not include ANOVA assumptions) of $\leq$0.85 [32]. The inter-observer cohort demonstrated high ICCs, 74.4% of which were $\geq$0.9. Such a high ICC value in the inter-observer setting is expected given the high Dice coefficients (>0.9) between the observers' contours. In comparison, the literature reports Dice coefficients ranging from 0.86 for non-small cell lung cancer (NSCLC) CT to 0.26 for mesothelioma CT; 91% of features to have ICC > 0.8 for NSCLC PET; and an average ICC of 0.77 for NSCLC CT-PET [8,33,34]. Inter-observer variability in MRI-radiomics has shown an average ICC of 0.85 for breast cancer and an ICC > 0.95 for all entropy features (only features examined) from diffusion-weighted MRI for cervical cancers [17,18]. The diagnostic–simulation test resulted in the fewest reproducible features, 14.1% of which have an ICC $\geq$ 0.75.

From this study, we draw three conclusions. Firstly, shape features demonstrated the highest repeatability and reproducibility in all tests. Shape features are commonly reported as highly reproducible in the literature, and were shown to be less sensitive to CT slice thickness and reconstruction parameters in a phantom study [35]. Further, shape features were found to be repeatable in test–retest of rectal cancer and NSCLC [7,8,13]. A recent systematic review, mostly based on CT studies, concluded that shape features were more reproducible than texture features, but that first-order features are better than both [6].

Secondly, the diagnostic–simulation cohort was designed to test features in a clinically relevant setting across different MR scanners. Reasonably, the number of reproducible features was fewer than the other two cohorts, likely due to differing image acquisition parameters. Of the 248 reproducible features in the diagnostic–simulation cohort, 92.3% was also reproducible in the

other two cohorts. Our findings are difficult to compare to the literature as specific features are uncommonly reported, especially given the sparse literature on MRI-based radiomics. Of the reproducible features identified in our study, coarseness has been reported as reproducible for breast cancer PET imaging [26]. Additionally, Fave et al. reported coarseness, gray length nonuniformity and run length nonuniformity as reproducible for NSCLC cone-beam CT [30]. Leijenaar et al. reported that GLCM and GLRLM were more reproducible than GLSZM, each of which encompasses at least one feature which appeared in our study as reproducible [36].

Thirdly, there is no substantial difference in feature stability between the original and filtered image domains. Wavelet and LoG-filtered images showed both better and worse reproducibility than the original images in the three cohorts tested in this study. Specifically with regard to the diagnostic–simulation cohort, this finding suggests that there is no filter or decomposition which overcame differences in acquisition parameters without losing the inherent tumor texture. Similarly, Schwier et al. demonstrated no significant improvement in reproducibility with a certain LoG-filter or wavelet decomposition [23]. Elsewhere, Timmeren et al. reported that wavelet features were less reproducible than the unfiltered image features in a test–retest scenario [8].

We acknowledge limitations in our study. This was a single-institutional retrospective study with a modest number of patients that may not be representative of other institutions or patients. However, our cohort size is very similar to those reported in the literature [14,37–39] and provides important results which highlight the pressing need for radiomic studies with larger cohorts. Additionally, this study focused on cervical cancer and its applicability to other tumor sites is unconfirmed. Despite the validation of the PyRadiomics platform, results may differ from other radiomic feature extraction platforms. The fixed bin method employed in this study is limited due to the increased number of bins once wavelet filters are applied. Further investigation on the effect of fixed bin width versus dynamic bin width is required. There was no bias field correction applied to the images in this study. The impact of field variation across the bore on feature reproducibility requires further study. Finally, although we used commonly reported cut-offs from the literature for ICC categories (0.75 and 0.9), these may not represent the ideal threshold for feature inclusion in prognostic models. While this study presents limitations, it has systematically evaluated MR-based radiomic reproducibility in three clinically applicable settings which has scarcely been done previously. Future work will involve analyses of the dependencies between radiomic features and clinical variables to better understand which radiomic features are the most appropriate for inclusion in prognostic models.

In conclusion, MRI-based radiomic features of cervical tumors were tested for their repeatability and reproducibility. Shape features emerged as the most reliable. The diagnostic–simulation resulted in the fewest reproducible features which highlights the importance of careful feature selection for radiomics generalizability. Further research in MRI-based radiomics is required to validate the use of reproducible features in prognostic models.

Conflicts of interest

None.

Acknowledgements

This work was supported by the Princess Margaret Cancer Center Radiation Medicine Program Radiogenomics/Radiomics Grant to Drs. Kathy Han and Michael Milosevic.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2019.03.001.

References

[1] Lambin P et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441–6.
[2] Scalco E, Rizzo G. Texture analysis of medical images for radiotherapy applications. Br J Radiol 2017;90:20160642.
[3] Kumar V et al. Radiomics: the process and the challenges. Magn Reson Imaging 2013;30:1234–48.
[4] O'Connor JPB, Rose CJ, Waterton JC, Carano RAD, Parker GJM, Jackson A. Imaging intratumor heterogeneity: Role in therapy response, resistance, and clinical outcome. Clin Cancer Res 2015;21:249–57.
[5] Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ. CT texture analysis: definitions, applications, biologic correlates, and challenges. RadioGraphics 2017;37:1483–503.
[6] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and reproducibility of radiomic features: A systematic review," Int. J. Radiat. Oncol., p. In press., 2018.
[7] Hu P et al. Reproducibility with repeat CT in radiomics study for rectal cancer. Oncotarget 2016;7:71440–6.
[8] Van Timmeren JE et al. Test–retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography 2016;2:361–5.
[9] T. Perrin et al., "Short-term reproducibility of radiomic features in liver parenchyma and liver malignancies on contrast-enhanced CT imaging," Abdominal Radiology, pp. 1–8, 2018.
[10] Talwar A et al. Pulmonary nodules: Assessing the imaging biomarkers of malignancy in a 'coffee-break'. Eur J Radiol 2018;101:82–6.
[11] Balagurunathan Y et al. Test–Retest Reproducibility Analysis of Lung CT Image Features. J Digit Imaging 2014;27:805–23.
[12] I. Shiri, H. Abdollahi, S. Shaysteh, and S. R. Mahdavi, "Test-Retest Reproducibility and Robustness Analysis of Recurrent Glioblastoma MRI Radiomics Texture Features," no. 5, p. e48035, 2017.
[13] Desseroit M-C et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. J Nucl Med 2017;58:406–11.
[14] Aerts HJWL et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.
[15] Haga A et al. Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. Radiol Phys Technol 2018;11:27–35.
[16] Huang Q et al. Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. J Med Imaging 2017;5:011005.
[17] Guan Y et al. Whole-lesion apparent diffusion coefficient-based entropy-related parameters for characterizing cervical cancers: initial findings. Acad Radiol 2016;23:1559–67.
[18] Saha A, Harowicz MR, Mazurowski MA. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. Med Phys 2018:3076–85.
[19] Lehmann TM, Gönner C, Spitzer K. Survey: Interpolation methods in medical image processing. IEEE Trans Med Imaging 1999;18:1049–75.
[20] Larue RTHM et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol. (Madr) 2017;56:1544–53.
[21] Van Griethuysen JJM et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–7.
[22] Zwanenburg A S, Leger S, Vallières M. Löck image biomarker standardization initiative. CancerData 2016.
[23] M. Schwier et al., "Repeatability of Selected Multiparametric Prostate MRI Radiomics Features," arXiv, 2018.
[24] Presotto L et al. PET textural features stability and pattern discrimination power for radiomics analysis: An 'ad-hoc' phantoms study. Phys. Medica 2018;50:66–74.
[25] Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med 2012.
[26] 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;55:414–22.
[27] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 2016;15:155–63.
[28] Zou KH et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004;11:178–89.
[29] Thada V, Jaglan V. Comparison of Jaccard, Dice, Cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. Int J Innov Eng Technol 2013;2:202–5.

[30] Fave X et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Med Phys 2016;42:6784–97.
[31] Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Methods Meas 2007;1:77–89.
[32] Berenguer R et al. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology 2018;288:172361.
[33] Pavic M et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol (Madr) 2018;57:1070–4.
[34] Parmar C et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9:e102107.
[35] Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L. Exploring variability in CT Characterization of tumors: a preliminary phantom study. Transl Oncol 2014;7:88–93.
[36] Leijenaar RTH et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol (Madr) 2013;52:1391–7.
[37] Balagurunathan Y et al. Reproducibility and prognosis of quantitative features extracted from CT images. Transl Oncol 2014;7:72–87.
[38] I. Shiri, H. Abdollahi, S. Shaysteh, and S. R. Mahdavi, ‘‘Test-Retest Reproducibility and Robustness Analysis of Recurrent Glioblastoma MRI Radiomics Texture Features,” Iran. J. Radiol., no. 5, p. e48035, 2017.
[39] Desseroit MC et al. Development of a nomogram combining clinical staging with 18F-FDG PET/CT image features in non-small-cell lung cancer stage I-III. Eur J Nucl Med Mol Imaging 2016;43:1477–85.