A multicenter study on radiomic features from $T_2$-weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics

Linda Bianchini$^1$ | João Santinha$^{2,3}$ | Nuno Loução$^4$ | Mário Figueiredo$^3$ | Francesca Botta$^5$ | Daniela Origgi$^5$ | Marta Cremonesi$^6$ | Enrico Cassano$^7$ | Nikolaos Papanikolaou$^2$ | Alessandro Lascialfari$^8$

$^1$Department of Physics, Università degli Studi di Milano and INSTM RU, Milan, Italy
$^2$Computational Clinical Imaging Group, Center for the Unknown (CCU), Champalimaud Foundation, Lisbon, Portugal
$^3$Instituto de Telecomunicações, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
$^4$Philips Healthcare, Lisbon, Portugal
$^5$Medical Physics Unit, IEO, European Institute of Oncology IRCSS, Milan, Italy
$^6$Radiation Research Unit, IEO, European Institute of Oncology IRCSS, Milan, Italy
$^7$Breast Imaging Division, IEO, European Institute of Oncology IRCSS, Milan, Italy
$^8$Department of Physics, Università degli Studi di Pavia and INSTM RU, Pavia, Italy

Correspondence
Linda Bianchini, Department of Physics, Università degli Studi di Milano and INSTM RU, Via Celoria 16 – 20133, Milan, Italy.
Email: linda.bianchini@unimi.it

Purpose: To investigate the repeatability and reproducibility of radiomic features extracted from MR images and provide a workflow to identify robust features.

Methods: $T_2$-weighted images of a pelvic phantom were acquired on three scanners of two manufacturers and two magnetic field strengths. The repeatability and reproducibility of features were assessed by the intraclass correlation coefficient and the concordance correlation coefficient, respectively, and by the within-subject coefficient of variation, considering repeated acquisitions with and without phantom repositioning, and with different scanner and acquisition parameters. The features showing intraclass correlation coefficient or concordance correlation coefficient $>0.9$ were selected, and their dependence on shape information (Spearman’s $\rho > 0.8$) analyzed. They were classified for their ability to distinguish textures, after shuffling voxel intensities of images.

Results: From 944 two-dimensional features, 79.9% to 96.4% showed excellent repeatability in fixed position across all scanners. A much lower range (11.2% to 85.4%) was obtained after phantom repositioning. Three-dimensional extraction did not improve repeatability performance. Excellent reproducibility between scanners was observed in 4.6% to 15.6% of the features, at fixed imaging parameters. In addition, 82.4% to 94.9% of the features showed excellent agreement when extracted from images acquired with echo times 5 ms apart, but decreased with increasing...
Computational advances have allowed the development of high-throughput analyses capable of extracting a large number of quantitative features from medical images. This process, called radiomics,\textsuperscript{1-3} has been studied as a means to provide imaging biomarkers that could assist in patients’ disease management, especially in oncology. The predictive models based on radiomic features showed preliminary, but promising results in guiding patients’ diagnosis, predicting response to treatment and prognosis, and in providing information on cancer genetics.\textsuperscript{4} Early clinical demonstrations of radiomics in MRI of the pelvis include prostate,\textsuperscript{5} rectal,\textsuperscript{6} and cervical\textsuperscript{7} cancer, with representative results in pelvic oncology summarized in Schick et al.\textsuperscript{8}

Despite the initial promising outcomes, some critical issues of the radiomic methodology are still unsolved. Radiomic features can be influenced by many factors,\textsuperscript{9} such as image acquisition parameters, magnetic field strength, reconstruction algorithms, intra- and interobserver segmentation variability, image processing, image artifacts, and choice of the software used for feature extraction. These factors can hinder the use of radiomic features as imaging biomarkers, requiring repeatability and reproducibility studies to assess their robustness to scanner/acquisition variability (e.g., scanner noise fluctuations, different acquisition parameters, different scanners), patient variability (anatomical and physiological deviations), and variability in image analysis (e.g., intra- and interobserver variability when using manual and semiautomatic segmentation).\textsuperscript{10-12}

The identification of repeatable and reproducible features can serve as an initial feature selection method that ensures the development of reliable models.\textsuperscript{11} Some studies have assessed the robustness of radiomic features in different modalities and applications, with the majority of the investigations performed in CT or positron emission tomography.\textsuperscript{12,13} The lack of standardization in the acquisition protocols and signal intensity scales, the high number of interplaying parameters affecting image quality and signal intensity, the lower spatial resolution (compared with CT), and a higher frequency of artifacts make radiomics analysis of MRI data very challenging. Gourtsoyianni et al.\textsuperscript{14} studied the repeatability of MRI features extracted from primary rectal cancer. Fiset et al.\textsuperscript{15} studied the repeatability and reproducibility of MRI-based radiomic features on cervical cancer. Recently, Schwier and colleagues\textsuperscript{12} investigated the repeatability of radiomic features on prostate tumors, assessing various normalization schemes, image pre-filtering, and bin widths for image discretization. Despite this thorough inspection, they did not find consistent improvements in repeatability across the different approaches. Scalco et al.\textsuperscript{16} found that the choice of image intensity normalization technique had a strong impact on the reproducibility of radiomic features, evaluated on $T_2$-weighted ($T_2$W) images of the pelvis.

Other studies were performed on phantoms, which offer more controlled experimental set-ups and the possibility of repeated scans. Most of the existing phantom studies investigated the reliability of radiomic features extracted from CT images. For example, Varghese et al.\textsuperscript{17} reported that CT-based texture features depend on the scanner and acquisition parameters. In another study by Baeßler et al.,\textsuperscript{18} the repeatability and reproducibility of MRI-based radiomic features under different matrix sizes were investigated for several imaging sequences using fruits as test objects. Yang et al.\textsuperscript{19} showed that radiomic features depend on the image acquisition process and reconstruction algorithm with experiments on a digital MR phantom. Furthermore, several clinical studies were conducted with variable acquisition parameters, known to affect the tissue contrast, whereas the influence of such changes in radiomic features was investigated to some extent by Mayerhoefer et al.\textsuperscript{20} and Chirra et al.\textsuperscript{21}

Although these studies provide some insight into the reliability of radiomic features in MRI, to the best of our knowledge, a systematic investigation on the repeatability and reproducibility of MRI-based radiomic features across clinical scanners/vendors and magnetic field strengths has not yet been performed.
In this study, we used a pelvic phantom designed explicitly for MRI radiomic purposes to assess the repeatability, with and without repositioning, as well as the reproducibility of two-dimensional (2D) and three-dimensional (3D) radiomic features extracted from T2W MR images acquired on three scanners from two vendors at two field strengths. First, we performed the experiments at fixed imaging parameters. Then, we preliminarily investigated the influence of the pulse repetition time (TR) and the echo time (TE) variability on the reproducibility of the radiomic features in one of the available scanners.

2 | METHODS

2.1 | Study design

In a clinical study, the database used for a radiomic investigation can include images acquired on different scanners and/or with different imaging protocols. We investigated the robustness of the radiomic features in some of these possible scenarios.

In this study (Figure 1), three scanners from two manufacturers at two magnetic field strengths were used to assess repeatability and reproducibility of radiomic features. Phantom images were acquired in a test–retest study to investigate the repeatability of the features on each scanner. The acquisitions were performed with a 2D T2W sequence, optimized on each scanner for pelvic imaging. Between the two repeated scans, the phantom was either kept fixed or repositioned, to evaluate also the repeatability after repositioning. Additionally, repeatability was investigated on 3D features extracted from images acquired using a 3D T2W sequence on scanner B.

The reproducibility of the features was evaluated in three scenarios: (1) reproducibility at fixed imaging parameters on two scanners of the same field strength, but from different manufacturers; (2) reproducibility at fixed imaging parameters on two scanners of the same manufacturer, but different field strengths (1.5T and 3T); and (3) reproducibility with varying TE and TR. Study 1 was performed on scanner A (1.5T Optima MR450W; GE Healthcare, Waukesha, WI) and scanner B (1.5T Ingenia; Philips Healthcare, Best, the Netherlands). The sequence parameters used for pelvic diagnostic imaging on scanner A were replicated on scanner B, and the phantom was imaged on both scanners. The replication of the same sequence in scanners of different vendors is not exact, as MR sequences are vendor-specific, and not all the parameters are accessible to users. However, considering that some radiomics retrospective studies made use of images acquired on different scanners, we have mimicked a multisensor scenario in close-to-ideal conditions (similar interscanner sequence parameters) to give useful indications for prospective studies. Study 2 was carried out on scanners B and C (3T Ingenia; Philips Healthcare), with a similar approach. Doubling the field strength, the MR signals are different, as they are influenced by the relaxation times T1 and T2.

**FIGURE 1** Study design schematic showing repeatability and reproducibility experiments and additional experiments executed to bring understanding to the repeatability and reproducibility results. *Clinical T2-weighted MRI sequence was shared between scanners B and C. FS, magnetic field strength*
T₂, which depend on the field strength. As a consequence, the radiomic features are expected to vary as well. However, this scenario could be part of a clinical radiomic study and would necessitate the assessment of the reproducibility of features between field strengths. Study 3 was conducted on scanner B, varying TE or TR in the range commonly used when imaging patients for pelvic examinations.

Besides, we evaluated the correlation between texture and shape features to identify whether the excellent performance in terms of repeatability and reproducibility might be ascribable to high correlation with shape information rather than robust quantification of a texture property.

2.2 | The phantom

A phantom designed for radiomic purposes was used for this study. The procedure to build the phantom was described in detail in Bianchini et al. Briefly, a pelvic-shaped container (NEMA IEC Body Phantom Set; Spectrum Corporation, Durham, NC) was used as the main phantom compartment, filled with a solution of 0.4mM MnCl₂ to reproduce the relaxation time T₂ of the muscle (40 to 65 ms at 1.5T) surrounding pelvic lesions. To study the repeatability on scanners B and C and the reproducibility between these scanners, the filling solution was replaced with oil (Spectrasyn 4 phantom oil; Philips Healthcare; T₁ = 230 ms and T₂ = 194 ms at 3T), which has a lower dielectric constant than water, allowing to avoid dielectric artifacts intensified at higher fields. Four inserts (Figure 2A), created by mixing polystyrene spheres of different diameters (1 to 8 mm) and 0.1% water solution of agar were embedded in the solution to provide a MR average signal and texture consistent with a set of representative pelvic tumors.

2.3 | Repeatability and reproducibility between scanners

2.3.1 | Images acquisition and segmentation

T₂W images of the phantom were acquired on scanners A, B, and C, using different pelvic imaging set-ups. The parameters of the MR sequences are listed in Table 1. The signal-to-noise ratio (SNR) was calculated according to method 4 described in the National Electrical Manufacturers Association standard (NEMA MS 1-2008 [R2014]).

To test the short-term repeatability of the radiomic features, the acquisition was repeated twice on each scanner without changing the set-up or moving the phantom. Then, the phantom was removed and repositioned, and the acquisition repeated. Sixteen cylindrical regions of interest (ROIs) of four sizes were drawn on the phantom inserts for the first acquisition images (Figure 2B) using 3D Slicer version 4.10.1 (http://www.slicer.org/). Each ROI was drawn on three consecutive slices, resulting in a cylindrical volume of different diameters (12, 24, 36, or 48 mm) and height equal to three times the slice thickness. The ROIs for the images acquired after phantom repositioning were obtained by applying to the original ROIs segmentation the rigid transformation that...
## Table 1: Scanner properties and imaging sequence parameters

| Scanner | Field strength (T) | Vendor | Experiment | Acquisition | TR (ms) | TE (ms) | SNR | Slice thickness (mm) | Spacing between slices (mm) | Pixel spacing (mm × mm) | FoV (mm × (mm) | Number of Averages |
|---------|-------------------|--------|------------|-------------|---------|---------|-----|---------------------|-----------------------------|-------------------------|----------------|-----------------|
| A       | 1.5               | GE     | Repeatability A/ reproducibility A vs. B | 2D          | 4763    | 109     | 70.95 | 5                   | 5.5                         | 0.6 × 0.6                | 320 × 320          | 1               |
|         |                   |        | Reproducibility A vs. B                | 2D          | 4700    | 110     | 24.58 | 5                   | 6                           | 0.6 × 0.6                | 320 × 320          | 1               |
|         |                   | Philips| Reproducibility B 3D                   | 3D          | 1050    | 160     | 369.92 | 1                   | 1                           | 1 × 1                    | 320 × 320          | 2               |
|         |                   |        | Reproducibility TE/TR                  | 2D          | 5000\(^a\) | 80 | 41.29 | 5 | 6 | 0.6 × 0.6 | 320 × 320 | 1 |
|         |                   |        |                                         |             | 85       | 36.56 |
|         |                   |        |                                         |             | 90       | 36.44 |
|         |                   |        |                                         |             | 95       | 34.19 |
|         |                   |        |                                         |             | 100      | 32.65 |
|         |                   |        |                                         |             | 105      | 32.26 |
|         |                   |        |                                         |             | 110      | 29.26 |
|         |                   |        |                                         |             | 115      | 28.92 |
|         |                   |        |                                         |             | 120      | 28.25 |
|         |                   |        | Repeatability B 2D/ reproducibility B vs. C | 2D          | 4405\(^b\) | 100 | 32.24 |
| B       | 1.5               | Philips| Reproducibility C/ reproducibility B vs. C | 2D          | 3750    | 90     | 377.23 | 5                   | 5                           | 0.6 × 0.6                | 340 × 340          | 1               |
|         |                   |        |                                         |             | 3750    | 90     | 452.91 | 5                   | 5                           | 0.6 × 0.6                | 340 × 340          | 1               |

Abbreviations: 2D, two dimensional; 3D, three-dimensional; FoV, field of view; SNR, signal-to-noise ratio; TE, echo time; TR, repetition time.

\(^a\)TR of 5000 ms corresponds to a manual selection of TR by the scanner operator.

\(^b\)TR of 4405 ms corresponds to an automatic selection of optimal TR in a range between 4000 ms and 6000 ms (Philips TR range operation mode).
allowed the alignment of the initial images with the images of the repositioned phantom.

2.3.2 Radiomic features extraction

PyRadiomics\textsuperscript{27} version 2.2.0 (https://github.com/Radiomics/pyradiomics/tree/2.2.0) was used to normalize the images and to extract the radiomic features, which included the following categories: Shape, First Order, GLCM (gray-level co-occurrence matrix), GLRLM (gray-level run length matrix), GLSZM (gray-level size zone matrix), NGTDM (neighboring gray-tone difference matrix), and GLDM (gray-level dependence matrix) from each of the 16 ROIs, on both original and filtered images (Laplacian of Gaussian—LoG-, wavelet, square, square root, logarithm, and exponential). We used a 6.0-mm σ in the LoG filter and one level for the wavelet decompositions. PyRadiomics preprocessing normalization was applied by setting the images mean signal intensity to 300 and standard deviation to 100. Assuming a normal distribution, the use of fixed bin number would add unknown effects and larger intensity resolution variability.\textsuperscript{12,29} The bin width was optimized for each extraction to obtain a number of bins in the range 30 to 130, which has shown good performance and reproducibility in the literature.\textsuperscript{30} (The parameter files are available at https://github.com/ReliabilityRadiomicsIEFOC/PhantomStudy. The list of features extracted from each category is reported in Supporting Information Table S1.)

2.3.3 Repeatability and reproducibility assessment

The repeatability of the radiomic features was quantified by computing the intraclass correlation coefficient (ICC).\textsuperscript{11,31,32} ICC (2,1), Equation (1), was calculated pairwise (between repeated acquisitions) for each radiomic feature to test repeatability, with and without phantom repositioning. The suffix (2,1) indicates the ICC form computed considering two-way random effects for absolute agreement and single rater/measurement.

\[
ICC(2,1) = \frac{MS_R - MS_E}{MS_R + (K-1)MS_E + \frac{K}{n}(MS_C - MS_E)}
\]

In Equation (1), \(MS_R\) corresponds to the mean square for ROIs, \(MS_E\) corresponds to the mean square for error, \(MS_C\) corresponds to the mean square for repeated measures, \(n\) is the number of ROIs, and \(K\) is the number of repeated acquisitions.

To test reproducibility, the concordance correlation coefficient (CCC)\textsuperscript{33} was calculated pairwise for the features extracted from the images acquired on two different scanners (scanner A versus B, scanner B versus C), with the same imaging sequence or acquired with different TE/TR. In Equation (2), \(\sigma_1^2\) and \(\sigma_2^2\) are the variances of a feature for each acquisition, \(\mu_1\) and \(\mu_2\) are the feature means, and \(\rho_{12}\) is the correlation coefficient between the acquisitions.

\[
CCC = \frac{2\sigma_1\sigma_2\rho_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}
\]

These metrics were computed using all 16 ROIs (four ROIs per insert). The features were classified into four groups based on the ICC or CCC values: excellent (ICC or CCC > 0.9), good (0.75 < ICC or CCC ≤ 0.9), moderate (0.5 < ICC or CCC ≤ 0.75), or poor (ICC or CCC ≤ 0.5) agreement. The most robust features for this particular scenario were identified by intersecting the sets of features showing both excellent repeatability and reproducibility.

The within-subject coefficient of variation (wCV)\textsuperscript{34,35} was also computed for both repeatability and reproducibility experiments to further extend utility of the findings to future studies. The wCV can be computed as follows:

\[
wCV = \frac{\sigma_e}{\mu_e}
\]

where \(\mu_e\) and \(\sigma_e\) are, respectively, estimated as \((nK)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}Y_{ik}^2\) and \(\sqrt{\sum_{i=1}^{n}\sum_{k=1}^{K}(Y_{ik} - \bar{Y}_i)^2}/n(K-1)\), with \(Y_{ik}\) corresponding to the measurement of sample \(i\) and replication \(k\), \(\bar{Y}_i = (1/K)\sum_{k=1}^{K}Y_{ik}\) and \(n\) has the same meaning as in Equation (1). The lower the wCV, the more repeatable the features will be. Repeatability and reproducibility as a function of the ROI size and the dependence on the contents of different inserts was also assessed by separating either ROI sizes or ROIs of different inserts.

2.4 Assessment of shape information in nonshape radiomic features

Besides repeatability and reproducibility issues, a recent study showed other vulnerabilities of radiomic features by assessing the performance of previously defined models on the same data sets, but where images had their voxel intensities randomly shuffled and found that three out of three nonshape features
constituting the model were actually capturing the tumor volume.36

Following this result, we performed an investigation on the nonshape features, aiming to identify the robust features because of their high correlation with shape features, but were not extracting the informative content their mathematical definition was expected to quantify. To this purpose, we proceeded with two consecutive analyses.

First, we identified the subset of nonshape features that were highly correlated with shape, through a pairwise-correlation analysis between shape and nonshape radiomic features, evaluated with Spearman’s correlation coefficient. Nonshape features showing Spearman’s correlation above 0.8 were considered as highly correlated with shape features. Other shape metrics besides volume37 were included because characteristics like maximum diameter,38 roundness,39 and spiculation,40,41 among others, are important oncological diagnostic and prognostic factors.

Second, three additional sets of images were created by shuffling the voxel intensities of each image corresponding to the first acquisitions performed on each scanner. The radiomic features were extracted from the shuffled intensities data sets, following the procedure described above for the segmentation and the calculation of features. An example of an original image and the corresponding image with randomly shuffled intensities is shown in Figure 3. ICC (2,1) was calculated between the original features and the corresponding features extracted from images that had their intensities randomly shuffled. Features showing ICC > 0.9 were considered texture uninformative because they could not distinguish the original image from the shuffled-intensities one in radiomic terms.

Intersecting the results of the two analyses, it was possible to identify a set of features to exclude a priori from an eventual radiomic model because they were both highly correlated with shape and not carrying texture information.

3 | RESULTS

3.1 | Repeatability and reproducibility between scanners

3.1.1 | Repeatability assessment

In the 2D assessment, 944 features were extracted (Figure 4A). Without repositioning, on scanner A, 910 (96.4%) features showed excellent, 29 (3.1%) good, 3 (0.3%) moderate, and 2 (0.2%) poor repeatability. Similar results were obtained on scanner B, with 869 (92.1%), 58 (6.1%), 8 (0.8%), and 9 (1.0%) features showing respectively excellent, good, moderate, and poor repeatability. The features extracted from the images acquired on scanner C showed less repeatability, with the number of features with excellent ICC decreasing to 754 (79.9%), and the number of features showing ICC ≤ 0.9 increasing to 190 (20.1%). When considering phantom repositioning, a consistent reduction of repeatability was evident across all the scanners. The numbers were 740 (78.4%), 138 (14.6%), 49 (5.2%), 17 (1.8%) (scanner A); 806 (85.4%), 76 (8.1%), 42 (4.4%), 20 (2.1%) (scanner B); and 106 (11.2%), 199 (21.1%), 215 (22.8%), 462 (44.9%) (scanner C) features showing excellent, good, moderate, and poor repeatability. With respect to wCV (Figure 4B), all scenarios, with the exception of scanner C with repositioning, had a majority of features with wCV ≤ 10%.
3D extraction, as shown in Supporting Information Table S2, did not increase the repeatability of the radiomic features.

Repeatability using wCV to assess possible trends between different sizes and inserts is shown in Supporting Information Figure S1.

3.1.2 Reproducibility assessment

Reproducibility was assessed both in terms of variations between the features extracted from scanners from different manufacturers and equal magnetic field strength and acquisition parameters (scanner A versus B; Figure 5A), and between features extracted from scanners from the same manufacturer, but different magnetic field strengths (scanner B versus C; Figure 5A). The CCC values are available at https://github.com/ReliabilityRadiomicsIEOFCAI/PhantomStudy.

The analysis of the features extracted from the images acquired on scanners A and B showed that 830 (87.9%) features had poor reproducibility. Only 43 (4.6%) features exhibited excellent reproducibility; 29 (3.1%) showed good and 42 (4.4%) moderate reproducibility.
In terms of reproducibility on scanners B (1.5T) and C (3T) evaluated with the CCC, 147 (15.6%), 295 (31.2%), 321 (34.0%), and 181 (19.2%) of them exhibited excellent, good, moderate, and poor reproducibility. A similar trend was found with the wCV evaluation, where more features with lower wCV were found between scanner B versus C than scanner A versus B (Figure 5B). Regarding the wCV evaluation, from Supporting Information Figure S2, similar behavior was found when comparing the reproducibility between scanner A versus B and scanner B versus C for differences between inserts and ROI sizes, where scanner B versus C yielded more features with wCV ≤ 10% than scanner A versus B, but no clear trend was found between inserts or ROI sizes.

The assessment of the reproducibility varying the TE showed that the majority of features (82.4% to 94.9%) had excellent reproducibility when extracted from images acquired with TEs 5 ms apart (Supporting Information Figure S3). The reproducibility decreased by increasing the TE interval. Changing the TR, 90.7% of the features exhibited excellent reproducibility.

The complete list of wCV of each feature for the different experiments and grouped by inserts and sizes are provided at https://github.com/ReliabilityRadiomicsIEOFC/PhantomStudy.
3.1.3 | Overall repeatability and reproducibility

Thirty-one (3.3%) features showed excellent repeatability (ICC > 0.9) and reproducibility (CCC > 0.9) across the scenarios studied (repeatability: with and without phantom repositioning; reproducibility: across manufacturers and field strengths). Apart from shape features, which are expected to be independent of the experiment settings (the reproducibility of the shape features shape_Flatness, shape_LastAxisLength, and shape_Sphericity—as 3D features—can be affected by differences in the through-place spacing), the 20 (2.1%) remaining features are listed in Table 2.

3.2 | Assessment of shape information in nonshape radiomic features

In this analysis, two subsets of radiomic features were identified: (1) nonshape features highly correlated with shape; (2) features showing a high correlation between their value when extracted from the original and the shuffled-intensities image.

| TABLE 2 | Features showing excellent repeatability and reproducibility |
|---------|---------------------------------------------------------------|
| log_sigma.6.mm.3D_firstorder_TotalEnergy | exponential_firstorder_Energy |
| exponential_firstorder_TotalEnergy | logarithm_grlm_RunLengthNonUniformity |
| original_grlm_RunLengthNonUniformity | square_firstorder_Energy |
| square_firstorder_TotalEnergy | squareroot_grlm_RunLengthNonUniformity |
| wavelet.HH_firstorder_Energy | wavelet.HH_firstorder_TotalEnergy |
| wavelet.HH_ngtdm_Coarseness | wavelet.HL_firstorder_Energy |
| wavelet.HL_firstorder_TotalEnergy | wavelet.HL_grlm_RunLengthNonUniformity |
| wavelet.HL_ngtdm_Coarseness | wavelet.LH_firstorder_Energy |
| wavelet.LH_firstorder_TotalEnergy | wavelet.LH_grlm_RunLengthNonUniformity |
| wavelet.LH_ngtdm_Coarseness | wavelet.LL_grlm_RunLengthNonUniformity |

List of the features showing excellent repeatability with and without phantom repositioning and excellent reproducibility across manufacturers and magnetic field strengths (Shape features, expected to be independent of the experiment settings, are not reported in this Table). Each feature is indicated in the form: ImageType_Class_FeatureName.

In set (1), out of 930 features, 155 (16.7%), 144 (15.5%), and 158 (17.0%) non-shape features were highly correlated with shape features in scanners A, B, and C, respectively. The discrepancy across scanners may be due to differences in SNR. A matrix showing the correlation between shape and non-shape features is reported, as example, in Supporting Information Figure S4. Ninety-nine (10.6%) features were shared among all scanners. Subset (2) had 27 (2.9%), 28 (3.0%), and 27 (2.9%) features for scanners A, B, and C, respectively.

When intersecting the features in subsets (1) and (2) common to all the scanners, 19 (2.0%) features were obtained. In this way, we identified the set of features nominally belonging to texture, but providing only shape information, listed in Table 3.

4 | DISCUSSION

Clinically meaningful radiomic signatures, providing high performance and generalizability, should be constructed using reliable features. Our study focused on several aspects that may influence the value of the radiomic features.

We investigated two types of 2D repeatability, offering measures on scanner-induced variations (without phantom repositioning) and repositioning-induced variations across scanners. Without repositioning, 92.1% to 96.4% of the features showed excellent repeatability in both 1.5T scanners. The percentage decreased to 79.9% when considering the images acquired at 3T, that may be because of the artifacts more frequently affecting images acquired at a higher field strength. When considering repositioning, a reduction in the number of features with excellent repeatability was observed: 78.4% of features showing excellent repeatability on scanner A, 85.4% on scanner B, and 11.2% on scanner C. Possible causes for the relatively small reduction observed with scanners A and B may be the rotational invariance of some features and a misalignment of the acquisition after repositioning. For scanner C, after visual inspection of the images, it was observed that the chemical shift artifact was considerably larger compared with the other scanners, as expected because it is proportional to the field strength. Additionally, as chemical shift artifacts occur in the frequency-encoding direction caused by the coexistence of water and lipid protons in a voxel, a slight change in the positioning of the phantom will be translated into a change in these artifacts. In terms of wCV (Figure 4B), most features showed a wCV ≤ 10% for all scanner and repeatability types (with and without repositioning), with the exception of scanner C with repositioning. Regarding differences between inserts and ROI sizes (Supplementary Information Figure S1), without repositioning, a larger percentage of features yielded a wCV ≤ 10%. The repeatability after repositioning based on the ROI sizes
for scanner C was below 25% for the different sizes. Further experiments and assessments of these dependencies are of major importance.

Comparing the 3D extraction from isotropic voxels and the 2D extraction, the 3D features did not improve repeatability with a possible explanation being the additional detail in the through-plane, leading to more unstable features after repositioning. In fact, a coarser spacing may not affect the performance of radiomic models and may even improve them as features become less susceptible to noise, repositioning, and other artifacts. These considerations justify the stability observed in the 2D features.

Our results showed a higher reproducibility when comparing two scanners of 1.5T and 3T from the same manufacturer (even though the relaxation times affecting the MR signal are dependent on the field strength), than comparing two scanners of 1.5T with same acquisition parameters, but from different vendors. Regarding reproducibility between different vendors at fixed imaging parameters, the images acquired on the two scanners exhibited different SNR values. Despite the basic principles of the sequence being the same, these differences can be caused by various user-independent sources. One factor is the type of systems, including the digital versus analog and the location of the analog-to-digital converter. As a matter of fact, Philips converts the signal to digital directly on the coil, whereas GE makes it on the magnet. Other factors consist of different preparation and calibration phases, including distinct power optimizations, frequency determination, shimming, and coils tune. Our study highlighted that even in the presence of unavoidable differences between images acquired with two scanners of different manufacturers, a subset of features turned out to be reproducible. As for the reproducibility between inserts or ROI sizes across scanners using wCV, only the percentage of features with wCV ≤ 1% seems to increase with ROI size, but additional experiments are required.

In this illustrative study, we considered a controlled pelvic-imaging scenario, resulting in only 31 radiomic features (3.3% of the total number of features extracted) showing excellent robustness in the considered settings. The scenarios of clinical studies may be different from the one herein considered, but an analogous procedure could be applied.

As demonstrated by Welch and colleagues, some features may exhibit dependencies on volume. However, these dependencies may be extended to other shape features that may also contain diagnostic and prognostic information. Therefore, it is possible that nonshape features may be stable based on the shape information they may contain. Nineteen nonshape features common to all three scanners appeared to contain only shape information or were heavily dependent on shape. These features showed both a high correlation with shape features and excellent agreement between features extracted from the original images and images with randomly shuffled intensities.

All the results discussed so far used images acquired with fixed-sequence parameters. However, the need for a large quantity of data to implement radiomic models may lead to the creation of inhomogeneous databases, as used in several clinical studies. This inhomogeneity might affect the values of radiomic features and the performance of models.

Our preliminary investigation on the influence of different TEs and TRs demonstrated that the majority of the features showed decreasing reproducibility as the TE interval increased. This corroborated our expectations for a T2W sequence, as a different TE changes the weighting in T2 and SNR, which will lead to variations of the T2 signal depending on the underlying tissue and pathophysiology. As texture depends on contrast, and because TE-induced differences are nonlinear, the dependence on contrast is not removed during image normalization. In fact, the increase of TE (from 80 to 120 ms) caused a progressive decrease in the SNR, which may have an impact on the radiomic feature extracted. As a consequence, the reproducibility of the features acquired with different TEs could be compromised. This idea was confirmed with the decreasing number of features showing excellent reproducibility (CCC

| TABLE 3 | Uninformative radiomic features |
|---------|-------------------------------|
| log.sigma.6.mm.3D_firstorder_Energy | log.sigma.6.mm.3D_firstorder_TotalEnergy |
| log.sigma.6.mm.3D_ngtdm_Coarseness | wavelet.LH_firstorder_Energy |
| wavelet.LH_firstorder_TotalEnergy | wavelet.LH_glrlm_RunLengthNonUniformity |
| wavelet.HL_firstorder_Energy | wavelet.HL_firstorder_TotalEnergy |
| wavelet.HL_glrlm_RunLengthNonUniformity | wavelet.HH_firstorder_Energy |
| wavelet.HH_firstorder_TotalEnergy | wavelet.HH_ngtdm_Coarseness |
| wavelet.HH_ngtdm_Coarseness | wavelet.LL_glrlm_RunLengthNonUniformity |
| square_firstorder_Energy | square_firstorder_TotalEnergy |
| square_glszm_GrayLevelNonUniformity | exponential_firstorder_TotalEnergy |

List of texture features showing, across all scanners, high correlation with shape features and uninformative about texture. Each feature is indicated in the form: ImageType_Class_FeatureName.
> 0.9) when extracted from images obtained with an increasing difference in TE. Similarly, 90.7% of the features showed excellent reproducibility between acquisitions with a TR of 5000 ms and 4405 ms. This suggests that radiomic studies should be conducted with standardized protocols or by making use of the features showing excellent reproducibility in the interval of TEs and TRs used, ensuring that the results are not associated with differences in the acquisition parameters.

Additional experiments to assess repeatability and reproducibility differences between feature classes, image filter types, and possible dependence on the field strength and parameters may also be performed to provide more information. As an example, a brief assessment to show the importance of this analysis is given in the Supporting Information Tables S3 and S4. From Table S3, wavelet LH provided more repeatable features than the original image with and without repositioning in 1.5T scanners (A and B) and for 3T (C) without reposition. As for Table S4, the GLSZM feature class provided less-repeatable features in all scenarios, with the exception of scanner C with repositioning, where GLCM provided a lower percentage of its total number of features.

Previous studies assessed the stability of radiomic features in different MRI settings. Mayerhoefer and colleagues20 investigated the sensitivity of texture features to acquisition parameters on T2W images acquired on a 3T scanner. They found that the imaging parameters influenced the feature value, with this influence increasing with spatial resolution. Although we did not compare different spatial resolutions, our study performed on a 1.5 T scanner extends their findings to a lower field strength, while making use of a phantom design for patient imaging, as we proved that the texture features values are dependent on the TE and TR.

In another study,21 prostate T2W images from different sites were used to assess reproducibility. The authors found that most of the Haralick features were reproducible in over 99% of cross-site comparisons. However, that study used different patient populations on each site and assessed nontumoral regions under the assumption that these should have a similar texture. Besides, part of their preprocessing involved image upsampling, in some cases by a factor of ~11 times in the image through-plane. Such choices make the comparison with our results difficult.

Our study has the following limitations. First, the use of a phantom cannot include all the effects that exist in clinical scenarios, such as patient motion, rectal/bladder filling, peristalsis, breathing, tissue diffusion and perfusion, and intrapatient tissue variability.

Undeniably, a phantom cannot be exhaustively representative of all tumors, so it cannot identify all the trustworthy and robust features for patient image analysis. On the other hand, contrary to the clinical reality, the use of a phantom allows repeating as many acquisitions as desired to compare results, and to assess the influence of many parameters. In particular, the use of our inhomogeneous phantom allowed the implementation of a procedure for identifying features that may not be trustworthy and robust. Furthermore, it allowed the assessment of the reproducibility issues caused by different system types, eg, digital versus analog, and coils used, which are translated into distinct image properties like the SNR that may have an impact on the values of radiomic features as well. Thus, the experiments conducted in this study provide excellent baseline assessments, under controlled environments, of the stability of features, and allow the avoidance of misleading results that could be derived without an accurate selection before the statistical analysis.

Another limitation was the use of data acquired on a restricted number of scanners and the focus on the investigation of the radiomics stability in T2W images as part of the clinical diagnostic protocol for pelvic imaging. Conversely, this can represent the starting point for further extension. Following these results, it might be interesting to make use of the phantom in an extended multicenter study using different clinically optimized sequences for a more comprehensive investigation.

All imaging data, radiomic files, parameter files for feature extraction, and analysis code are provided so that researchers can extend this analysis and tune these experiments to their needs.

5 | CONCLUSION

In conclusion, this study investigated the robustness of radiomic features extracted from T1W images of a pelvic phantom created for MR radiomic analyses. Our investigation quantified the stability and quality of radiomic features in different MRI settings, enlightening important issues towards robust models.

Based on a workflow designed to test repeatability and reproducibility, features showing the highest performance were identified. Importantly, many of these repeatable and reproducible features turned out to be inadequate for radiomic analysis, being noninformative or affected by the image-acquisition process. We recommend to apply this, or similar, procedure to strengthen each clinical radiomic study.

ACKNOWLEDGMENTS

Professor Alessandro Lascialfari and Professor Nikolaos Papanikolaou contributed equally to this work. The authors thank the Nuclear Medicine Departments at Champalimaud Clinical Center—Champalimaud Foundation and at European Institute of Oncology for the availability to use the NEMA IEC Body Phantom Set.
DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available in GitHub at https://github.com/ReliabilityRadiomicsIEOFCPHantomStudy. Additionally, imaging data are available at https://central.xnat.org/data/projects/ReliabilityRad.

ORCID
Linda Bianchini https://orcid.org/0000-0002-0081-9417
João Santinha https://orcid.org/0000-0001-8174-2943

TWITTER
Linda Bianchini @lindabianchin13

REFERENCES
1. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441-446.
2. Aerts HJWL, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:1-9.
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. Radiology. 2016;278:563-577.
4. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys Med Biol. 2016;61:R150-R166.
5. Sun YU, Reynolds HM, Parameswaran B, et al. Multiparametric MRI and radiomics in prostate cancer: A review. Australas Phys Eng Sci Med. 2019;42:3-25.
6. Zhou X, Yi Y, Liu Z, et al. Radiomics-based pretherapeutic prediction of non-response to neoadjuvant therapy in locally advanced rectal cancer. Ann Surg Oncol. 2019;26:1676-1684.
7. Li Z, Li H, Wang S, et al. MR-based radiomics nomogram of cervical cancer in prediction of the Lymph-vascular space invasion preoperatively. J Magn Reson Imaging. 2018;49:1-7.
8. Schick U, Lucia F, Dissaux G, et al. MRI-derived radiomics: Methodology and clinical applications in the field of pelvic oncology. Br J Radiol. 2019;92:20190105.
9. Traverso A, Wei L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: A systematic review. Int J Radiat Oncol Biol Phys. 2018;102:1143-1158.
10. Abramson RG, Arlinghaus LR, Dula AN, et al. MRI biomarkers in oncology clinical trials. Magn Reson Imaging Clin. 2016;24:11-29.
11. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment. Stat Methods Med Res. 2015;24:26-27.
12. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of multiparametric prostate MRI radiomics features. Sci Rep. 2019;9:1-16.
13. Mackin D, Faye X, Zhang L, et al. Measuring computed tomography scanner variability of radiomics features. Invest Radiol. 2015;50:757-765.
14. Gourtsoyianni S, Doundou G, Prezzi D, et al. Primary rectal cancer: Repeatability of global and local regional MR imaging texture. Radiology. 2017;284:1-10.
15. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. Radiother Oncol. 2019;135:107-114.
16. Scalco E, Belfatto A, Mastropietro A, et al. T2w-MRI signal normalization affects radiomics features reproducibility. Med Phys. 2020;47:1680-1691.
17. Vargheese BA, Hwang D, Cen SY, et al. Reliability of CT-based texture features: Phantom study. J Appl Clin Med Phys. 2019;20:155-163.
18. Baeßler B, Weiss K, Pinto dos Santos D. Robustness and reproducibility of radiomics in magnetic resonance imaging. Invest Radiol. 2019;54:221-228.
19. Yang F, Dogan N, Stoyanova R, Ford JC. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground truth. Phys Medica. 2018;50:26-36.
20. Mayerhoefer ME, Szomolanyi P, Jirak D, Materka A, Trattnig S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study. Med Phys. 2009;36:1236-1243.
21. Chirra P, Leo P, Yim M, et al. Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI. J Med Imaging. 2019;6:1.
22. Bianchini L, Botta F, Origi D, et al. PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis. Phys Medica. 2020;71:71-81.
23. ExxonMobil. SpectraSyn 4—Polyalphaolefin (PAO) Fluid. 2019. [Online]. Available: http://exxonmobilchemical.ides.com/en-US/ds244622/SpectraSynTM4.aspx?I=30156&U=0. [Accessed: 02-Jul-2020].
24. Moldoveanu SC, David V. Mobile phases and their properties. Essentials in Modern HPLC Separations. 2013:363-447.
25. National Electrical Manufacturers Association. Determination of Signal-to-Noise Ratio (SNR) in Diagnostic Magnetic Resonance Imaging. NEMA MS 1-2008 (R2014). Arlington, VA: NEMA; 2015.
26. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging. 2012;30:1323-1341.
27. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77:e104-e107.
28. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One. 2019;14:1-14.
29. Leijenaar RTH, Nalbantov G, Carvalho S, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: The need for standardized methodology in tumor texture analysis. Sci Rep. 2015;5:1-10.
30. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med. 2011;52:369-378.
31. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155-163.
32. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology of multiparametric prostate MRI radiomics features. Magn Reson Imaging Clin. 2019;77:e104-e107.
33. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45:255-268.
34. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative imaging biomarkers: A review of statistical methods for computer algorithm comparisons. *Stat Methods Med Res*. 2015;24:68-106.

35. Obuchowski NA, Barnhart HX, Buckler AJ, et al. Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example. *Stat Methods Med Res*. 2015;24:107-140.

36. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother Oncol*. 2019;130:2-9.

37. Zhang J, Gold KA, Lin HY, et al. Relationship between tumor size and survival in Non-Small-Cell Lung Cancer (NSCLC): An analysis of the Surveillance, Epidemiology, and End Results (SEER) registry. *J Thorac Oncol*. 2015;10:682-690.

38. Fukuhara H, Kume H, Suzuki M, et al. Maximum tumor diameter: A simple independent predictor for biochemical recurrence after radical prostatectomy. *Prostate Cancer Prostatic Dis*. 2010;13:244-247.

39. Lee SI, Oliva E, Hahn PF, Russell AH. Malignant tumors of the female pelvic floor: Imaging features that determine therapy: Pictorial review. *Am J Roentgenol*. 2011;196(3 SUPPL.):15-23.

40. Snoeckx A, Reyntiens P, Desbuquoit D, et al. Evaluation of the solitary pulmonary nodule: Size matters, but do not ignore the power of morphology. *Insights Imaging*. 2018;9:73-86.

41. Liu S, Wu X-D, Xu W-J, Lin Q, Liu X-J, Li Y. Is there a correlation between the presence of a spiculated mass on mammogram and luminal a subtype breast cancer? *Korean J Radiol*. 2016;17:846.

42. Rizzo S, Botta F, Raimondi S, et al. Radiomics: The facts and the challenges of image analysis. *Eur Radiol Exp*. 2018;2:36.

43. Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging*. 2020;20:1-10.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

**FIGURE S1** Repeatability of radiomic features by inserts (left column—A.1, B.1, C.1) and by ROI sizes (right column—A.2, B.2, C.2) using wCV. Note: A.1 and A.2, B.1 and B.2, C.1 and C.2 correspond to scanners A, B and C, respectively

**FIGURE S2** Reproducibility of radiomic features by inserts (A) and by ROI sizes (B) using wCV. Note: (A vs. B)—represent the reproducibility between scanners A and B; (B vs. C)—represent the reproducibility between scanners B and C

**FIGURE S3** Reproducibility of radiomic features with varying TEs. Reproducibility was assessed for all possible combinations of TEs between 80 ms and 120 ms in 5 ms intervals and the four reproducibility levels: (A) Percentage of features with excellent reproducibility; (B) Percentage of features with good reproducibility; (C) Percentage of features with moderate reproducibility; (D) Percentage of features with poor reproducibility. In each plot, a box represents the percentage of features showing CCC in a certain range (specific for each plot) when comparing the feature values extracted from images obtained with a combination of different TEs. The combination of TEs (in red) for a specific box can be read on of the corresponding row and column

**FIGURE S4** Correlation matrix of shape and non-shape radiomic features for the original images acquired on scanner A

**TABLE S1** List of radiomic features per category extracted with the package PyRadiomics. Shape features are extracted from the original images only, whilst the features of the other categories are extracted from both the original and filtered images. The definition of the features listed in this table is available at https://pyradiomics.readthedocs.io/en/latest/features.html

**TABLE S2** Repeatable 3D features. List of 3D radiomic features showing excellent repeatability in more than 80% of the image filters

**TABLE S3** Repeatability by filter type. Number of features per filter types showing excellent repeatability without and with (repos) repositioning for scanners A, B, and C

**TABLE S4** Repeatability by feature class. Absolute and relative number (shown inside parenthesis, %) of features per feature class showing excellent repeatability without and with (repos) repositioning for scanners A, B, and C

**How to cite this article:** Bianchini L, Santinha J, Loução N, et al. A multicenter study on radiomic features from T2-weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. *Magn Reson Med*. 2021;85:1713–1726. [https://doi.org/10.1002/mrm.28521](https://doi.org/10.1002/mrm.28521)