

# SCIENTIFIC REPORTS

OPEN

## Repeatability of Multiparametric Prostate MRI Radiomics Features

Michael Schwier<sup>1,2</sup>, Joost van Griethuysen<sup>3</sup>, Mark G. Vangel<sup>2,4</sup>, Steve Pieper<sup>5</sup>, Sharon Peled<sup>1,2</sup>, Clare Tempany<sup>1,2</sup>, Hugo J. W. L. Aerts<sup>2,6</sup>, Ron Kikinis<sup>1,2,7,8</sup>, Fiona M. Fennessy<sup>1,2,6</sup> & Andriy Fedorov<sup>1,2</sup>

Received: 29 August 2018

Accepted: 12 June 2019

Published online: 01 July 2019

In this study we assessed the repeatability of radiomics features on small prostate tumors using test-retest Multiparametric Magnetic Resonance Imaging (mpMRI). The premise of radiomics is that quantitative image-based features can serve as biomarkers for detecting and characterizing disease. For such biomarkers to be useful, repeatability is a basic requirement, meaning its value must remain stable between two scans, if the conditions remain stable. We investigated repeatability of radiomics features under various preprocessing and extraction configurations including various image normalization schemes, different image pre-filtering, and different bin widths for image discretization. Although we found many radiomics features and preprocessing combinations with high repeatability (Intraclass Correlation Coefficient > 0.85), our results indicate that overall the repeatability is highly sensitive to the processing parameters. Neither image normalization, using a variety of approaches, nor the use of pre-filtering options resulted in consistent improvements in repeatability. We urge caution when interpreting radiomics features and advise paying close attention to the processing configuration details of reported results. Furthermore, we advocate reporting all processing details in radiomics studies and strongly recommend the use of open source implementations.

The field of Radiomics is concerned with the extraction of quantitative imaging features to convert images into a large scale mineable data<sup>1</sup>. Lambin *et al.*<sup>2</sup> state the Radiomics hypothesis “that advanced image analysis on conventional and novel medical imaging could capture additional information not currently used, and [...] that genomic and proteomics patterns can be expressed in terms of macroscopic image-based features.” The prognostic and discriminative power of radiomics features has been explored in cancer imaging with promising results<sup>3–13</sup> (including tumor locations prostate, lung, head and neck, brain, breast, glioblastoma, etc.).

Prostate cancer is one of the emerging applications with a strong need for improved characterization of the disease using imaging, as is evident from the ongoing efforts to standardize acquisition and reporting of the imaging findings<sup>14,15</sup>. Multiparametric MRI (mpMRI) is a well-established clinical tool used effectively for cancer detection, characterization, treatment planning and response assessment. However, applications of quantitative analysis of mpMRI are very limited in the clinic.

The generally accepted standard of care is to use the Prostate Imaging Reporting and Data System (PI-RADS v2)<sup>15</sup>, which establishes the guidelines for performance of and qualitative interpretation of mpMRI. In the research applications, most of the studies investigating quantitative analysis of mpMRI utilize basic imaging-derived features such as lesion volume<sup>16</sup>, summary statistics of the Apparent Diffusion Coefficient (ADC)<sup>17</sup> and pharmacokinetic maps estimated from DCE MRI<sup>18</sup>. More recently, early results suggest that radiomics may have a role in differentiating non-cancerous benign prostate tissue from cancer, as well as grading prostate cancer<sup>19</sup>. Fehr *et al.*<sup>4</sup> combined a set of first and second order texture features computed in ADC and T2-weighted (T2w) in an automatic Gleason pattern classification algorithm. Wibmer *et al.*<sup>5</sup> also demonstrated that second order texture features on ADC and T2w may differentiate between cancer and normal tissue. They also found a correlation between ADC Entropy and Gleason score, but no correlation between T2w texture features and Gleason score. In another study Peng *et al.*<sup>6</sup> found that “[t]he combination of 10th percentile ADC, average ADC, and T2-weighted skewness with CAD is promising in the differentiation of prostate cancer from normal tissue. ADC image features and  $K^{\text{trans}}$  moderately correlate with GS”. In a very recent development, Bonekamp

<sup>1</sup>Brigham and Women’s Hospital, Boston, MA, USA. <sup>2</sup>Harvard Medical School, Boston, MA, USA. <sup>3</sup>Netherlands Cancer Institute Maastricht University, Amsterdam, Netherlands. <sup>4</sup>Massachusetts General Hospital, Charlestown, MA, USA. <sup>5</sup>Isomics, Inc, Cambridge, MA, USA. <sup>6</sup>Dana-Farber Cancer Institute, Boston, MA, USA. <sup>7</sup>Fraunhofer MEVIS, Bremen, Germany. <sup>8</sup>Mathematics/Computer Science Faculty, University of Bremen, Bremen, Germany. Correspondence and requests for materials should be addressed to A.F. (email: [andriy.fedorov@gmail.com](mailto:andriy.fedorov@gmail.com))

*et al.* concluded that the mean ADC has performance comparable to that of radiomic-based machine learning in identifying biopsy-confirmed clinically significant lesions<sup>20</sup>. As such, no unequivocal recommendation exists on what, if any, radiomics features can be recommended for PCa characterization.

The premise of radiomics is that quantitative image features can serve as a biomarker characterizing the disease, or allowing prediction of response and thus providing decision support for patient management. To reliably derive conclusions based on any biomarker, a basic requirement is that its value must remain stable between the two measurements, if the conditions remain stable<sup>21–24</sup>. In particular, this means that a biomarker must be stable under usual scanner noise and normal anatomical or physiological deviations. Investigation of biomarker repeatability is a fundamental component of its technical (assay) validation: one of the consensus recommendations for imaging biomarker (IB) validation and qualification in cancer studies<sup>25</sup> specifically states that “IB precision must be demonstrated early in IB development through single-centre repeatability studies, or few-site reproducibility studies,” which is exactly the aim of our study. We use the term “repeatability” for this attribute of a radiomics feature (others refer to it as “reproducibility<sup>21,22,26,27</sup>”, or “stability<sup>3,23,24,28</sup>”). Good repeatability is a necessary, but not a sufficient condition for a high predictive power of a feature, meaning that if a feature has a high predictive power, its repeatability must be good. If a feature has a low repeatability, its predictive power must be low, too. But if a feature has a good repeatability, we cannot conclude anything about its predictive power. Gudmundsson *et al.*<sup>24</sup> demonstrated this aspect in their evaluations of feature importance and stability on different physiological time series.

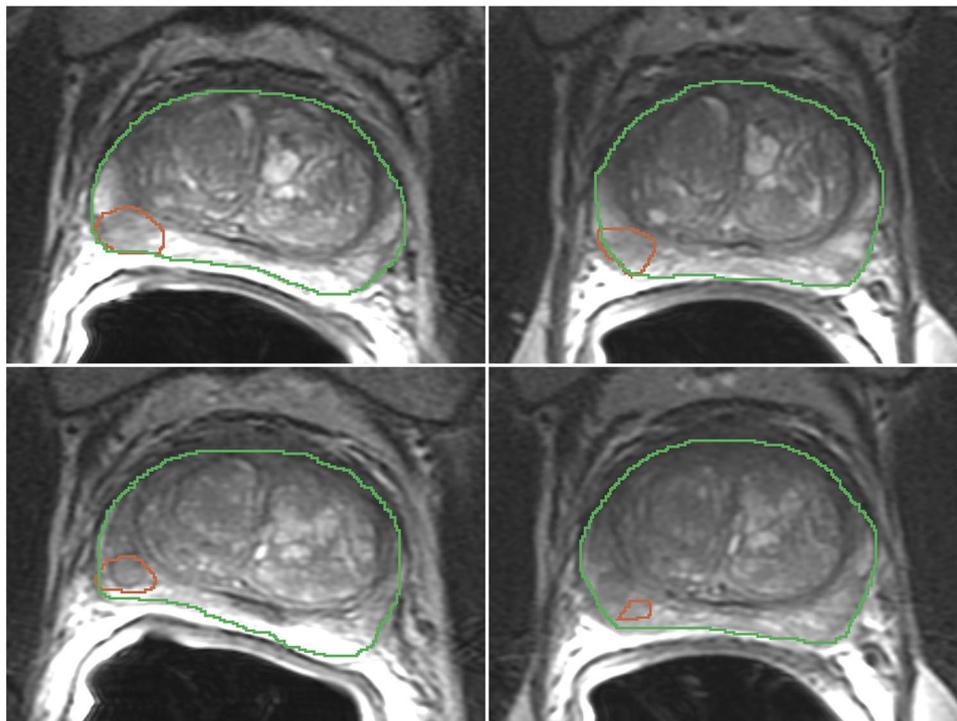
Considering the repeatability of features is therefore a good measure for pre-selecting features for a classification task, given a large amount of features to select from. Such a pre-selection is necessary since hundreds of feature sets are available for consideration in medical imaging<sup>29</sup>. This number multiplies if we consider different parameters, filters and preprocessing combinations.

Repeatability of imaging biomarkers has been investigated for a number of imaging modalities and applications. Zhao *et al.*<sup>30</sup> investigated repeatability of manual and computer aided diameter and volume measurements of lung lesions on CT test-retest. Others looked into the repeatability of MRI specific measures like per-voxel ADC<sup>31</sup> or quantitative parameters maps in T1 and T2\*- weighted images<sup>32</sup>. Bologna *et al.*<sup>28</sup> presented a method for assessing repeatability and predictive power of radiomics features without using test-retest data or multiple segmentations per case. Several studies have investigated a large set of radiomics features on CT lung cancer cases<sup>3,21,22,26,27,33</sup>. All of them found a large number of features with a good repeatability. Zhao *et al.*<sup>21</sup> found that many features are reproducible even under different CT reconstruction settings, but they also mention that repeatability of texture features is particularly susceptible to varying pre-processing schemes. Hu *et al.*<sup>34</sup> report 252 of 775 texture features with high repeatability on CT rectal cancer cases and that the influence of various filters on texture features is small. Another set of repeatability studies was conducted on non-small cell lung cancer cases<sup>23,35,36</sup>. While Leijenaar *et al.*<sup>23</sup> and van Velden *et al.*<sup>35</sup> report overall good repeatability, Desseroit *et al.*<sup>36</sup> mention a critical aspect: “repeatability [...] varied greatly among metrics” and “depended strongly on the quantization step, with different optimal choices for each modality”. Critical issues were also raised by Emaminejad *et al.*<sup>37</sup> and Chalkidou *et al.*<sup>38</sup>. Emaminejad *et al.*<sup>37</sup> investigated various factors influencing texture feature calculation (Entropy) and discovered that they introduce substantial variation. Chalkidou *et al.*<sup>38</sup> reviewed 15 studies and “found insufficient evidence to support a relationship between PET or CT texture features and patient survival”.

Based on these results, we feel that further emphasis on repeatability is needed in the radiomics literature. At the same time, radiomics analysis is fraught with complexities in identifying the optimal analysis parameters. As an example, we did not identify a consistent recommendation on how pre-filtering should be performed in PCa MRI radiomics characterization<sup>4–6</sup>. Depending on the specific study, image normalization (scaling and shifting) was applied only for texture features<sup>5</sup>, for all features<sup>4</sup>, or not used at all<sup>6</sup>. 3D computation of texture features is only mentioned in one study<sup>5</sup>, while others<sup>4,6</sup> do not specify whether their computations were done in 2D or 3D. Overall, the description of the preprocessing often lacks details to allow for exact reproduction of the calculations.

Furthermore, most of the existing studies investigating radiomics feature repeatability focus on features extracted from CT. Radiomics analysis of MRI data poses significant challenges due to lack of signal normalization, more common acquisition artifacts, and lower spatial resolution. In their comprehensive review radiomics paper of 2016 Yip *et al.*<sup>39</sup> state that “the repeatability of MR-based radiomic features has not been investigated” and that “[u]nderstanding the stability of MR-based radiomic features between test and re-test scans can help identifying reliable features for radiomic applications, and thus would be a valuable future study”. We still observe this gap in the present understanding of the feature repeatability applied to MRI analysis in general, and in prostate cancer imaging specifically.

In this study we assess the repeatability of radiomics features using a publicly available dataset<sup>40</sup> of small prostate tumors in multiparametric prostate MR images (mpMRI). We consider all features implemented in the open source *pyradiomics* package<sup>41</sup>, which are, for the most part, implemented according to consensus definitions of the Imaging Biomarkers Standardization Initiative (IBSI)<sup>29</sup>. We investigate various factors likely to influence the repeatability of features, such as image normalization, 2D/3D texture computation, discretization with different bin widths, and image pre-filtering. Furthermore, repeatability also depends on the accuracy of tumor segmentation at both time points but we did not fully investigate the impact of that factor in this study. In our reporting we focus on disclosing all configuration details and make our implementation and data available. We note that we do not report results for all of the observations (all combinations of image types, regions, choices of normalization, pre-filtering and feature sets). Instead, our goal is to summarize findings of most relevance. This study is an extension of our previous work, where we evaluated the repeatability of volume and apparent diffusion coefficient (ADC)<sup>42</sup>. Those basic imaging features are widely recognized as valuable markers of prostate cancer<sup>16,17,43</sup>.



**Figure 1.** Segmentation of the Tumor ROI (red) and Whole Gland (green) in T2w images for case 7 baseline (left) and follow-up (right) scans. Each slice pair was selected to match location. Top and bottom show consecutive slices going through the same tumor. The individual time points were segmented manually by the domain expert blinded to the other time point. Upon the review of the images corresponding to the time points side by side, it becomes apparent that the location of the segmented region is not consistent for the slice shown in the bottom row.

## Methods

**Image data and segmentations.** This study used a previously published, publicly available prostate mpMRI test-retest dataset<sup>40</sup> composed of fifteen treatment-naïve men with biopsy-confirmed ( $n = 11$ ) (using a sextant biopsy technique) or suspected ( $n = 4$ ) prostate cancer (PCa). After providing informed consent, all 15 patients underwent a second MR within two weeks after the first MR, without any interim treatment<sup>42</sup>. From the MRI exam we used the T2-weighted axial (T2w) images (TR 3350–5109 ms, TE 84–107 ms, FOV 140–200 mm), and the Apparent Diffusion Coefficient (ADC) maps derived from Diffusion-weighted MRI ( $b = 1400 \text{ s/mm}^2$ , TR 2500–8150 ms, TE 76–80 ms, FOV 160–280 mm). See Fig. 1 for an illustration of one of the images used in this study.

A radiologist with 10+ years of experience in prostate mpMRI reviewed all of the images for the individual MR studies using 3D Slicer software (<http://slicer.org>)<sup>44</sup>. Regions of interest (ROIs) annotated included the suspected tumor, entire peripheral zone of the prostate gland, and the entire prostate gland identified in the baseline and follow-up T2w and ADC images. Image annotation utilized a visualization protocol whereas all of the individual images for a given patient and time point combination were shown to the reader for a single time point. However, the individual timepoints were randomized so that while annotating a given study the reader was blinded to the other time point for the same patient. Notably, all resulting tumor ROIs used for calculating the features were smaller than 0.8 ml. Details relating to the acquisition and annotation of the dataset are described elsewhere<sup>40,42</sup>.

Upon review of the data, one of the subjects (case 1) was excluded from the analysis of ADC features, since one time point was deemed of poor image quality due to noise and observed distortions. As a result, analysis of repeatability of the radiomics features for the ADC images was conducted using the data from the remaining 14 subjects.

Stability of the features is affected by the consistency of the segmentation of the region of interest between the baseline and repeat scans. In absence of the ground truth, we cannot evaluate the absolute accuracy of the manual segmentation performed by the radiologist.

**Feature extraction.** Features were extracted for all ROIs using *pyradiomics*, presented earlier in<sup>41</sup>. We extracted features from five feature classes: First Order, Shape, Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), and Gray Level Run Length Matrix (GLRLM) features (throughout the text, whenever they correspond to implementations in *pyradiomics*, feature classes are capitalized, and feature names as well as preprocessing filters are capitalized and denoted in italics). All features of each of these classes were extracted with the following exceptions: *Compactness1* and *Compactness2*, as well as *SphericalDisproportion* from

Shape features were excluded because they are directly correlated to *Sphericity* (based on the definition of the feature, as discussed in the documentation of *pyradiomics*). *Flatness* and *LeastAxis* from Shape features were excluded because some tumor ROIs were only defined on one slice and these features do not yield useful values for non-3D objects. *SumAverage* was excluded because it is directly correlated with *JointAverage*. *Homogeneity1* and *Homogeneity2* were disabled because they are directly correlated to *InverseDifferenceMoment*.

*Pyradiomics* allows preprocessing of (applying filtering to) the original image before feature extraction and offers the following options<sup>41</sup>: *Original* - leave the image unchanged, *LoG* - apply a Laplacian of Gaussian filter, *Wavelet* - apply wavelet filters (all combinations of high- and low-pass filters on each image dimension), *Square* -  $(\text{image intensity})^2$ , *Square Root* -  $\sqrt{|\text{image intensity}|}$ , *Logarithm* -  $\log(|\text{image intensity}| + 1)$ , and *Exponential* -  $e^{(\text{image intensity})}$ . The last four filters also scale the values back to the original image range and restore negative sign if original was negative. For the *LoG* preprocessing we choose kernel sizes (sigmas) 1.0, 2.0, 3.0, 4.0, and 5.0 mm. Each feature was computed with each of the above mentioned preprocessing steps separately. We note that image pre-processing prior to feature calculation is currently not covered by the IBSI radiomics standardization initiative<sup>29</sup>. Here we consider all standard pre-processing approaches implemented in *pyradiomics*, which include pre-processing filters (i.e., *Wavelet* and *LoG*) that have been shown to result in highly predictive feature sets<sup>3,45</sup>.

Bias correction was applied to all T2w images to compensate for intensity non-uniformities using N4 Bias Correction approach<sup>46</sup> implemented in 3D Slicer<sup>44</sup>.

MR image intensity is usually relative and not directly comparable between images. To reduce this effect we applied intensity normalization. To test the effect of normalization on feature repeatability we included features computed with and without normalization. Whole-image normalization was performed by scaling and shifting the values of the whole image to a mean signal value of 300 and a standard deviation of 100. This means we would expect most of the values in the range of 0–600, assuming normal distribution of the intensities within the image. We also considered normalization based on a biologically comparable reference tissue region, which is assumed to be stable across time points and patients. For this we selected an ROI in a muscle region. The utility of normalization to the muscle reference region has been demonstrated before in prostate imaging studies (e.g., see Huang *et al.*<sup>47</sup>). Compared to the PZ tissue, we expect the muscle region to be more homogeneous, which we believe makes the muscle-reference normalization more robust. The shifting and scaling factor for all image voxels was determined such that the mean signal in the reference ROI changed to 100 and the standard deviation to 10. The smaller range and lower mean for the muscle-reference normalization is due to the fact that the reference region represents only a small portion of the whole image intensity range and that muscle tissue has low intensity. This way the intensities of the whole image maintain a reasonable range after muscle-reference normalization.

Computation of texture features requires discretization (binning) of the image intensities into a limited number of grey levels. This can be done using either fixed number of bins, or the fixed bin size (see section 2.7 of the IBSI guidelines v6<sup>29</sup>). We used the latter discretization approach, as implemented in *pyradiomics*. Leijenaar *et al.*<sup>48</sup> argue why it is imperative to use a fixed bin width and not a fixed number of bins for discretisation. Additionally, according to Tixier *et al.*<sup>49</sup> the total number of bins for texture feature computation should be between 8 and 128. Considering our value range of 0–600 after whole-image normalization, we should select bin widths not smaller than 5. However, for non-normalized images as well as muscle-reference normalization the intensity ranges vary and can be much larger. Hence, we selected bin widths 10, 15, 20, and 40 for performing our experiments, which should result in the number of bins below 128 for intensity ranges of up to 5120. We note that image binning as implemented in *pyradiomics* is based only on the intensities that are within the region of interest.

Texture features are computed as various statistics over specific matrices (e.g., co-occurrence matrix for GLCM). The dimensionality of the texture matrix defines the neighborhood (2D vs 3D) over which feature calculation is performed (see the *pyradiomics* documentation for more details). Since the choice between 2D- and 3D-based calculation of texture features is not obvious, and no comparisons of the two were done before, we included the comparison of stability of the two approaches in our study.

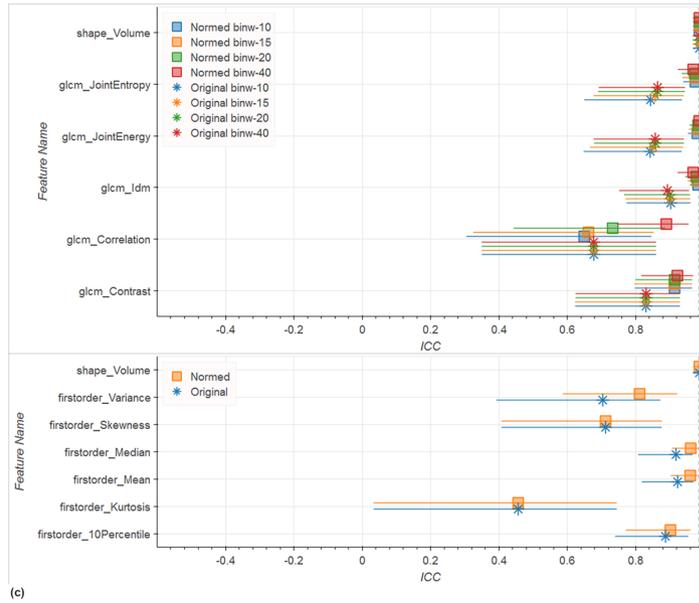
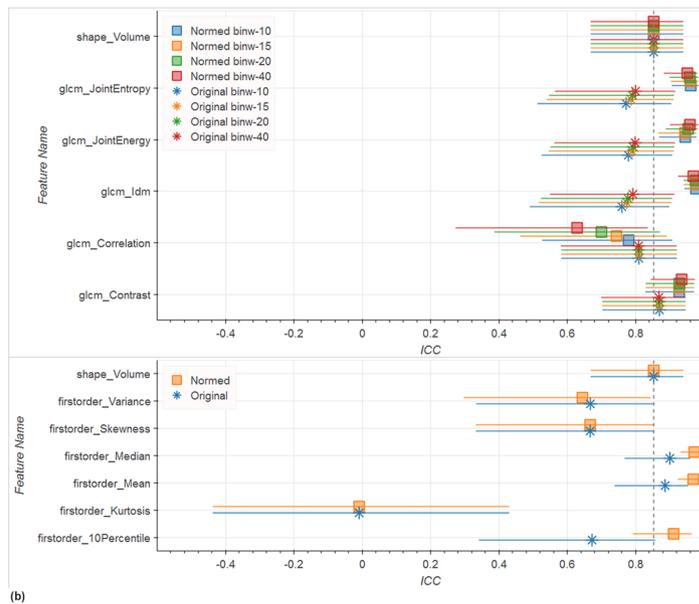
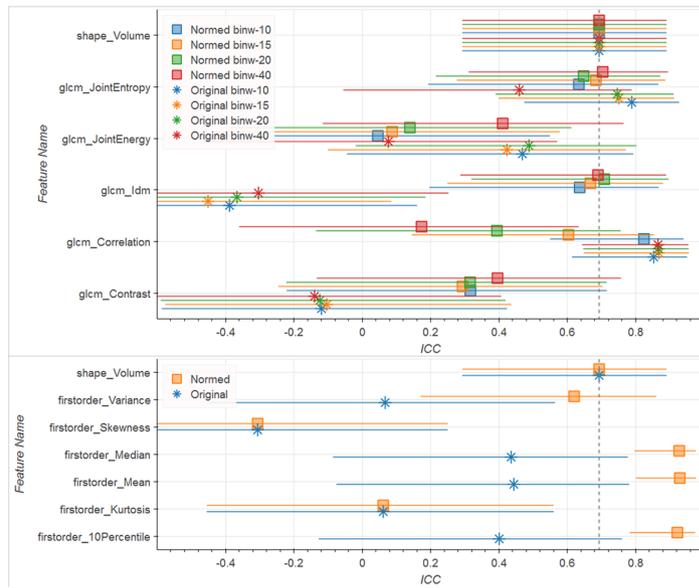
Default settings were used for all other configuration parameters of *pyradiomics* feature extraction (see <http://pyradiomics.readthedocs.io> for further information).

**Measure of repeatability.** As a measure of repeatability we report the intraclass correlation coefficient ICC(1,1)<sup>50</sup>. All scans were acquired on the same scanner for the baseline and repeat images, and were annotated by the same radiologist. Annotation was done by a single radiologist completely blinded with respect to subject. Therefore, the variability due to annotation is part of the within-subject variability. The two sources of variability included in the ICC are between-subject and between-scans within the subject. Since there is no systematic difference between the first and second scans, we model both between-subject and within-subject sources of variability as independent random effects and ICC(1,1) is therefore appropriate. The ICC considers the variation between repeated scans on the same subject in relation to the total variability in the population<sup>51</sup>. For our test-retest scenario with two time points it is defined as follows:

$$ICC(1, 1) = \frac{BMS - WMS}{BMS + WMS},$$

where BMS is the between-subjects mean squares and WMS the within-subjects mean square<sup>50</sup>. Hence BMS is an estimate for the variance between patients in our study and WMS an estimate for the variance over repeated measurements on the same patient.

The ICC is invariant with respect to linear scaling and shifting. This is a necessary property for using it to compare repeatability of features which operate in different unit and scale spaces. Since fixed thresholds for interpreting the ICC are problematic (see our Discussion section and e.g. Raunig *et al.*<sup>51</sup>), ICCs of different radiomics features should be compared to a reference within the study population. We use the *Volume* ICC as such



**Figure 2.** ICC and corresponding 0.9 confidence interval for selected features computed on different ROIs on ADC images. Texture features are computed in 2D. Colors represent the bin width for the texture computations, glyph shape represents if the image was whole-image-normalized or not normalized. No filtering was applied to the image. The dashed line indicates the reference *Volume* ICC. Results show that whole-image normalization tends to improve ICCs while bin width has only marginal influence. (a) Tumor ROI (top: texture features, bottom: first order features); (b) Peripheral Zone (top: texture features, bottom: first order features); (c) Whole Gland (top: texture features, bottom: first order features).

reference. Tumor volume is an important quantitative measure characterizing PCa, which has been investigated earlier<sup>16</sup>, and evaluation of its repeatability in this specific dataset has already been presented in our earlier work<sup>42</sup>.

**Evaluation approach.** We start by evaluating the repeatability of a small subset of features, which were shown by others to perform well in PCa mpMRI radiomics-style analyses<sup>4–6</sup>. Specifically, this initial step of the evaluation focused on the following Intensity and GLCM radiomics features: *Volume*, *Entropy*, *Energy*, *Idm* (Inverse Difference Moment), *Correlation*, *Contrast*, *Variance*, *Skewness*, *Median*, *Mean*, *Kurtosis*, and *10Percentile* (10th percentile of intensity distribution). Using this relatively small dataset, we first explore the effects of whole-image normalization and intensity bin size on the repeatability of those features for image types ADC and T2w. Based on the observations on this reduced feature set, we aim to identify preprocessing and feature extraction parameters that lead to improved repeatability, and continue with the evaluation of the complete radiomics feature set using the selected processing options. Since it is challenging to look into hundreds of features individually, in this phase we focus either on summarizing statistics over all features, or on a selection of the top 3 best performing features per feature group.

The large amount of data generated by our extraction also does not allow us to explore all aspects in detail in this paper. Further details can be found in an extended preprint<sup>52</sup>. The preprint also covers additional analyses not considered in this paper, such as an investigation of registration to mitigate the effect of potentially inconsistent segmentations between timepoints as well as an evaluation of repeatability on Subtraction images (pixel-wise difference between the early post-contrast image phase and the pre-contrast phase of the Dynamic Contrast Enhanced MRI acquisition). Furthermore, the source code accompanying this manuscript includes additional investigations.

**Ethical approval and informed consent.** The analysis conducted in this study was performed on the publicly available de-identified human subject data released earlier within the QIN-PROSTATE-Repeatability collection on The Cancer Imaging Archive (TCIA)<sup>40</sup>. Since the data is de-identified, was available prior to the presented study, and no identifiable information was used in the present analysis, our study qualifies for Exemption 4 under the “Basic HHS Policy for Protection of Human Research Subjects”, see <https://humansubjects.nih.gov/human-specimens-cell-lines-data>.

## Results

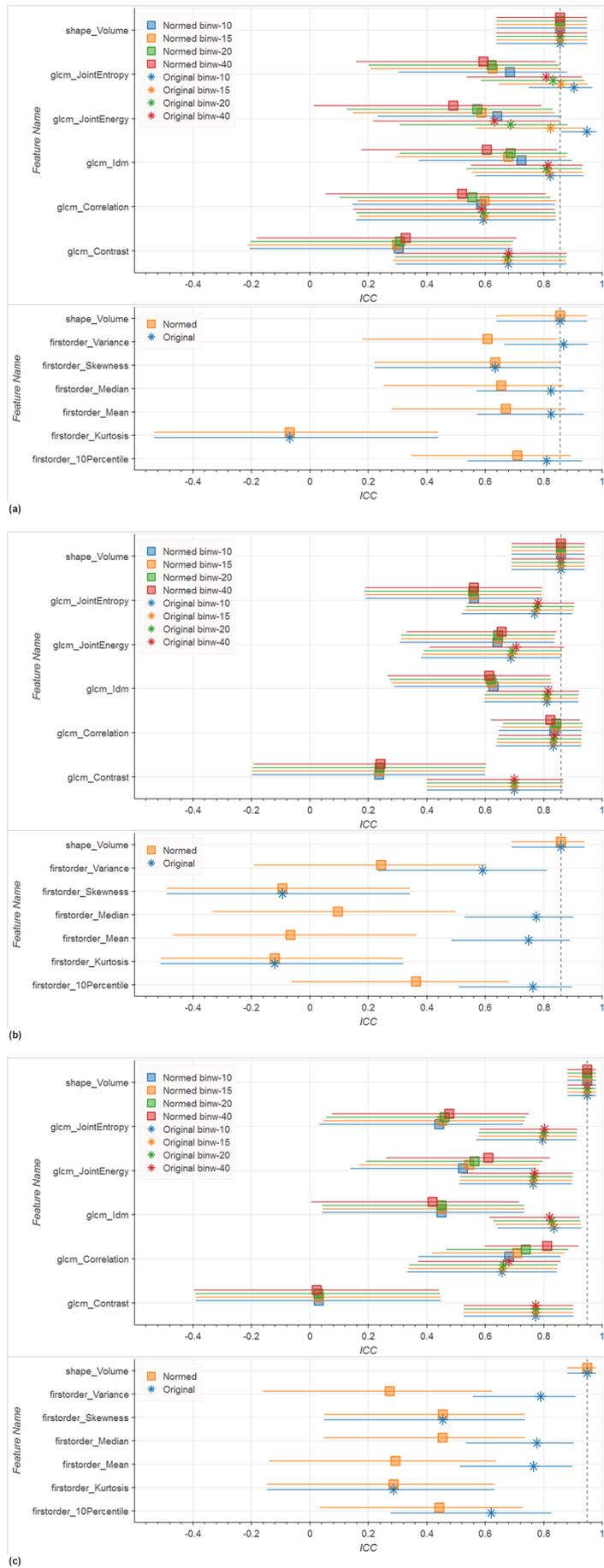
To investigate the repeatability of radiomics features for small prostate tumors in mpMRI we analyzed a large set of features under various preprocessing combinations. As there are numerous combinations of image types, regions, choices of normalization, pre-filtering and feature sets, we discuss the most interesting and applicable findings in our results. Additional results can be found in the extended preprint<sup>52</sup> of this paper.

**Selected features on whole-image-normalized vs non-normalized images.** Our evaluation of a small subset of features that were shown by others to perform well in PCa mpMRI analysis yielded the following results. For the Tumor ROI in whole-image-normalized ADC images (see Fig. 2a) *Entropy*, *Idm* (Inverse Difference Moment), *Correlation*, *Median*, *Mean*, and *10Percentile* (10th percentile of the intensity distribution) reach ICCs equal or better than *Volume* (ADC *Volume* ICC = 0.7). In the Peripheral Zone in whole-image-normalized ADC images (see Fig. 2b) *Entropy*, *Energy*, *Idm*, *Contrast*, *Median*, *Mean*, and *10Percentile* reach ICCs around 0.91 or higher, performing better than *Volume* (ADC *Volume* ICC = 0.85). Looking at the ADC ICCs in the Whole Gland (see Fig. 2c) we observe that no feature reaches the 0.99 ICC of *Volume*. Whole-image normalization leads to improved ICC in ADC images in most cases (see Fig. 2a–c). Notable exceptions are *Variance* in the Peripheral Zone as well as *Entropy*, *Energy*, and *Correlation* in the Tumor ROI. For these exceptions the difference to the ICC of the corresponding whole-image-normalized feature was always smaller than 0.1. Note that *Skewness* (measure of asymmetry of the distribution about the mean) and *Kurtosis* (measure of peakedness of the distribution) are by definition not influenced by whole-image normalization. However, they also never reach the reference *Volume* ICC.

On T2w images neither in the Peripheral Zone, nor in the Whole Gland does an ICC of any feature reach the *Volume* ICC (see Fig. 3b,c), which is 0.86 for the Peripheral Zone and 0.95 for the Whole Gland. Only in the Tumor ROI *Entropy*, *Energy*, and *Variance* reach an ICC higher than the reference T2w *Volume* ICC of 0.86 (see Fig. 3c). Contrary to ADC images, whole-image normalization leads to lower ICCs in T2w images. The only exception is *Correlation*, which has a higher ICC for the Whole Gland when normalized.

In the following we report further results based on whole-image-normalized ADC and non-normalized T2w images only, since results in this section showed better overall repeatability under these configurations.

**Influence of different bin-widths.** Figures 2 and 3 indicate that different bin widths do not result in very strong variations of the ICC for most features of the selected subset. The kernel density estimation (KDE)<sup>53,54</sup> plot in Fig. 4a,b support this observation. It shows the distribution of the maximum difference between highest



**Figure 3.** ICC and corresponding 0.9 confidence interval for selected features computed on different ROIs on T2w images. Texture features are computed in 2D. Colors represent the bin width for the texture computations, glyph shape represents if the image was whole-image-normalized or not normalized. No filtering was applied to the image. The dashed line indicates the reference *Volume* ICC. Results show that omitting whole-image normalization tends to improve ICCs while bin width has only marginal influence. (a) Tumor ROI (top: texture features, bottom: first order features); (b) Peripheral Zone (top: texture features, bottom: first order features); (c) Whole Gland (top: texture features, bottom: first order features).

and lowest ICC per feature depending on bin width for all GLCM features and all pre-filtering options. For the majority of features the maximum difference is around 0.2 or lower.

Another insight into the influence of bin width can be gleaned from Fig. 4c,d. We ranked the ICC for each feature depending on the bin width and plotted the rank distribution. We can see that for the lowest and highest bin widths, the best and worst ranks are appearing about equally often. Bin width 15 and 20 cover the middle ranks.

In the following we report all results based on bin width 15, since the results in this section indicate that this yields a reasonable average estimate of the feature repeatability.

**Top 3 features per feature group.** Given all possible combinations of pre-filtering options, we selected the 3 most repeatable features for each of the feature classes implemented in *pyradiomics* (namely Shape, First Order, GLCM, GLSZM and GLRLM). We then investigated whether any specific pre-filtering approach consistently resulted in improved repeatability of these selected features. For these top 3 features Fig. 5 illustrates the range of ICCs under all pre-filtering options for the Tumor ROI and Peripheral Zone in T2w images. We observe that other shape features have a better repeatability than Volume. They also, by definition, are not influenced by any pre-filtering. The ICCs of other features are scattered over a wide range. In the Tumor ROI (Fig. 5a) no single filter appears to result in consistently more stable features. However, several filters consistently result in ICCs below Volume ICC (e.g., *Logarithm* and *Exponential*). In the Peripheral Zone, only a few *Wavelet* filters yield ICCs above the reference for all top 3 features with the exception of GLCM *ClusterProminence*, for which also *Logarithm* filtering reaches a higher ICC. On the low end particularly the *Exponential* filter performs consistently weak.

Figure 6 illustrates the same top 3 analysis on ADC images for Tumor ROI and Peripheral Zone. In the Tumor ROI the ICCs are scattered over a wide range again (see Fig. 6a). Even though many filters are related to high ICCs on several features, no consistent trend can be observed. A few features have a strong tendency towards yielding high ICCs, like *LoG* sigma 3.0 mm, but we can always find an exception. In the Peripheral Zone most filters are associated with ICCs above the reference for almost all features (see Fig. 6b). Also, we observe that the spread of ICCs depending on the pre-filter is much smaller with a few exceptions (e.g., *GLSZM SmallAreaEmphasis*, or *Firstorder 90Percentile*).

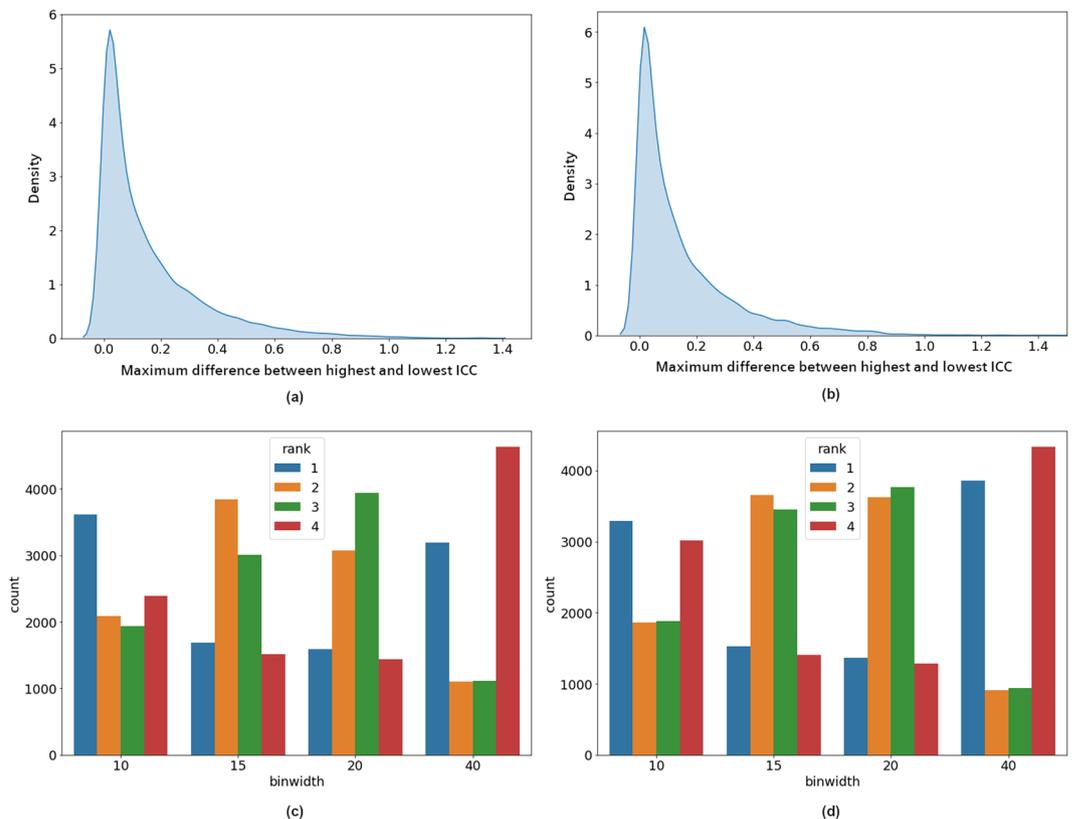
**Overview of pre-filter performance over all features.** For an overview of the influence of pre-filtering options on repeatability of all features we considered all features and pre-filter combinations on T2w and ADC images in the Tumor ROI and Peripheral Zone. We selected all of those combinations that had ICC above the corresponding Volume ICC. The plots in Figs 7 and 8 show how often a particular filter appears among those for Tumor ROI and Peripheral Zone in T2w and ADC images. On T2w Tumor ROI *Wavelet-HH*, *Wavelet-LH*, and smaller sigma *LoG* filters are most prominent. Also the other *Wavelet* filters as well as the original image (no filter applied) often lead to high ICC values. On T2w Peripheral Zone only *Wavelet-HH*, and *Wavelet-LH* stand out. In the Tumor ROI on ADC images *Original* and *LoG* filters perform well, while among the *Wavelet* filters *Wavelet-LH* and *Wavelet-LL* stand out. Among the Single Pixel filters *Square* shows a strong performance. For ADC Peripheral Zone the filters corresponding to high ICCs are more equally distributed. The only exceptions are *Exponential* and *Logarithm*, which both are less often associated with good repeatability. Overall, no filter is consistently associated with high ICC values - neither on T2w nor on ADC images (see Figs 7 and 8 for details).

**Normalization against a muscle reference region.** Normalization against a muscle reference region on T2w results in a strong decline of repeatability for most of the selected features (see Fig. 9a). We also again selected the top 3 most repeatable features per feature class for the original T2w as well as the T2w that was normalized against a muscle reference region. Figure 9b shows that some features remain stable but repeatability of most original T2w top 3 features strongly declines on muscle-reference-normalized T2w and vice versa.

## Discussion

Despite the small sample size and iterative exploratory approach to conducting the evaluation, our study of radiomics feature repeatability resulted in a large number of observations. To a degree, this was caused by our attempts to find patterns and to be able to explain the results and make specific recommendations. In the following we discuss our interpretation of the main findings of the study. Additional findings are reported in our pre-print<sup>52</sup> of this report, including further preprocessing options as well as an analysis of Subtraction images (SUB). We excluded SUB images from this study, because results did not provide additional insights. Furthermore, they are derived images, obtained by subtracting time frames of the DCE acquisition, hence introducing another source of uncertainty (frames to be subtracted are selected manually, there is a possibility of motion between the time frames). Finally DCE provides limited complementary value in prostate mpMRI<sup>55</sup>.

In this study we focus on repeatability under the premise of a realistic clinical patient re-evaluation, i.e., patients returning for a follow-up scan. Hence differences in patient position as well as the state and layout of organs (e.g., rectal/bladder filling) are expected. A radiomics feature useful for clinical practice must be robust to



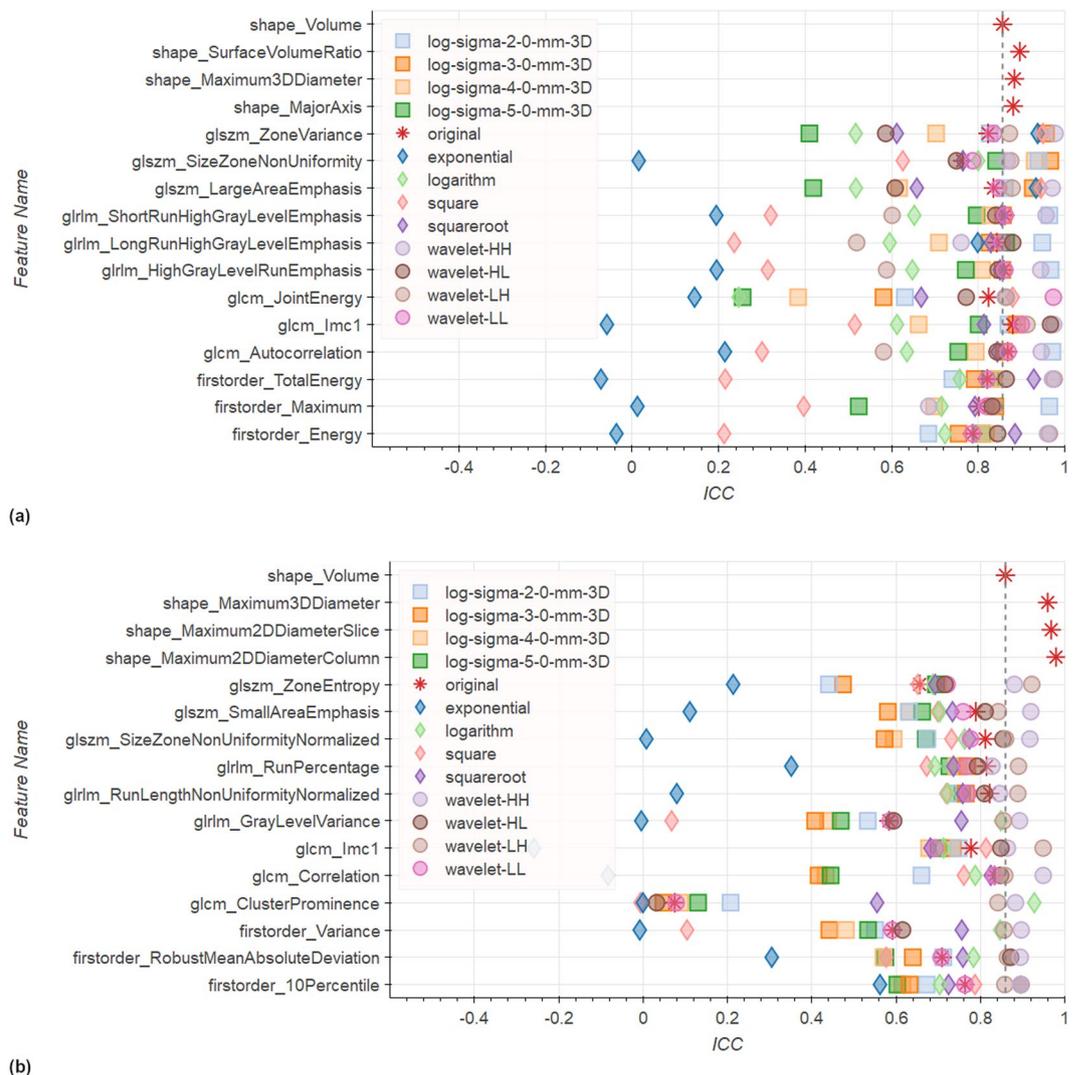
**Figure 4.** (a,b) Kernel density estimation plots for the ICC range over four bin widths, for all combinations of texture features and pre-filtering options on (a) ADC and (b) T2w images. For most feature and pre-filtering combinations, this range is small. (c,d) Histograms of the ICC ranks for all texture feature and pre-filter combinations for each bin width on (c) ADC and (d) T2w images. Bin widths 10 and 40 are very often associated with the best ICCs but similarly often with the worst ICCs. Bin widths 15 and 20 are more often associated with ICCs in the middle ranks. Hence bin widths 15 or 20 are a better choice to estimate the representative average repeatability of features.

such deviations. Another type of repeatability study would be to consider consecutive scans of the same patient without repositioning. This would focus only on repeatability variations induced by the scanner only, and was studied elsewhere (e.g., see the study by Sadinski *et al.*<sup>31</sup>). To the best of our knowledge, there is no publicly available dataset corresponding to the latter design, and therefore repeatability of radiomics features under such setting was not feasible in our study.

Among the features reported in the literature for prostate cancer analysis<sup>4–6</sup> in T2w and ADC MR images, we observed good repeatability ( $ICC \geq \text{Volume ICC}$ ) for *JointEntropy*, *Idm*, *Median*, and *Mean*. However, we could not confirm good repeatability of other features calculated over Tumor ROIs. Notable examples include *Kurtosis* and *Contrast*, which in all cases are underperforming compared to the ICC of *Volume*, in some cases reaching values close to 0. Furthermore, even features with good repeatability showed these only under specific preprocessing configurations. On our data, whole-image normalization led to improved ICCs for most features calculated from ADC images, while for T2w images whole-image normalization did not result in ICC improvements.

The observation that whole-image normalization improved repeatability for ADC is not completely unexpected. Although ADC is a quantitative measure that is supposed to be consistent across exams and platforms, variability of around 5% was observed even under perfect conditions for a temperature controlled phantom<sup>56</sup>. Several prominent studies in prostate MRI radiomics, e.g., see Fehr *et al.*<sup>4</sup> and Kwon *et al.*<sup>57</sup>, and elsewhere<sup>58</sup> apply normalization while analyzing ADC images. Others, like Bonekamp *et al.*<sup>20</sup>, do not apply normalization to ADC images. In practice, there does not seem to be a consensus on the use of normalization for ADC maps.

A possible explanation of the lack of improvement in T2w features' reproducibility extracted from the whole-image-normalized images could be the following. One of the main reasons we want to normalize MRI images is because we know different scanners and/or protocols yield different images. In the present dataset though, the same scanner and protocol is used. Therefore, the pre-normalization scans may be more comparable than may be expected in a clinical setting. Coupled with the fact that whole-image normalization uses the entire image to determine the mean and standard deviation, one might argue that in this case, noise over the entire image (e.g., due to different positioning, different bowel filling, etc.) might lower the comparability of intensities in the target region (prostate). The same reasoning might explain the ambiguous results for applying muscle-reference normalization to T2w images, resulting in improvements for some but decline in repeatability for other features.



**Figure 5.** Top 3 features for each feature group by ICC in T2w images for (a) the Tumor ROI and (b) the Peripheral Zone. Results in these figures illustrate that ICCs are spread over a wide range depending. Also some filters have a consistently low performance. However, no filter consistently performs above reference. We can also see that some shape filters have a higher repeatability than *Volume*. See Supplementary Dataset 1 for confidence intervals for the ICCs in this figure.

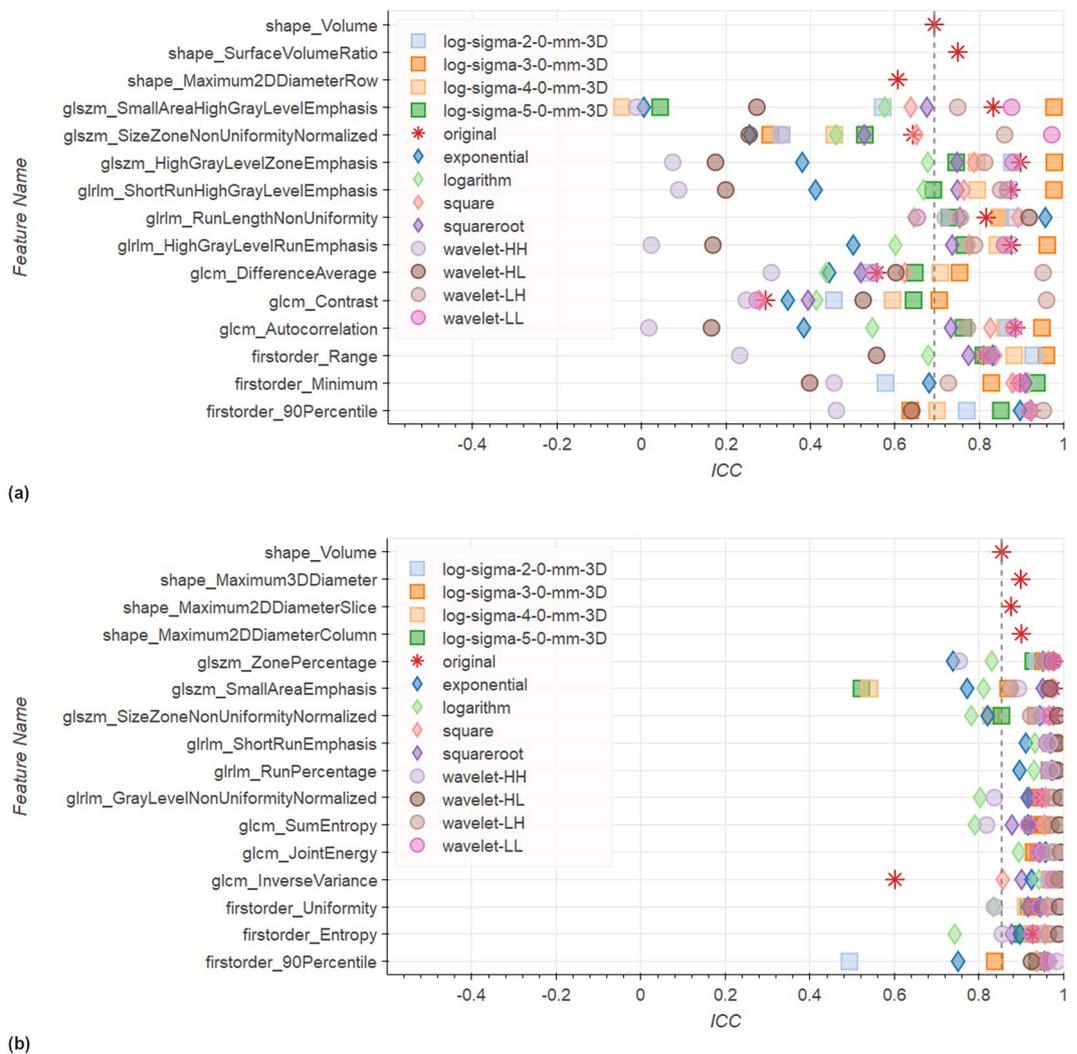
As we saw in Fig. 4, even when the bin width is selected within the recommended limits<sup>48,49</sup>, it still has an influence on the repeatability, although it is not too strong in most cases. Nevertheless, the differences we observed lead us to advice to evaluate the influence of bin width in any new study.

Our investigation of pre-filtering options and parameterization of the texture feature computation revealed further challenges in extracting repeatable radiomics signatures. We found that the use of pre-filtering introduces even more variation in the ICCs per feature across the various pre-filtering options (see Figs 5 and 6). Notable exception is the relative stability of features in the Peripheral Zone ROI for the ADC images (Fig. 6b). Furthermore, our analysis of how often each filter is related to an ICC above *Volume* reference (Figs 7 and 8) again reveals that we cannot suggest a filter that consistently improves repeatability.

Nevertheless, we can observe certain trends for filters that predominantly relate to below-reference ICCs. On T2w images these are all single pixel filters as well as large sigma *LoG*. On ADC images the results are less consistent. *Logarithm* and *Exponential* have low repeatability in the Peripheral Zone, while for the Tumor ROI *Logarithm*, *Wavelet-HH*, and *Wavelet-HL* have the weakest performance. Hence, based on our results, we recommend to leave these filters out for the designated image types and structures.

However, in any case there are still many other filters which yield high ICCs, but not consistently enough to be able to pick a few for general recommendation. Even if we narrow down to a specific image and structure, we cannot single out a small set of pre-processing configurations which consistently result in improved ICCs. The results are too diverse. Depending on the feature, different preprocessing options yield the best repeatability.

There is an exception, however. Our results show several shape features (e.g., *SurfaceVolumeRatio* and different uni-dimensional diameter measurements) with a better repeatability than *Volume*. By definition these are also



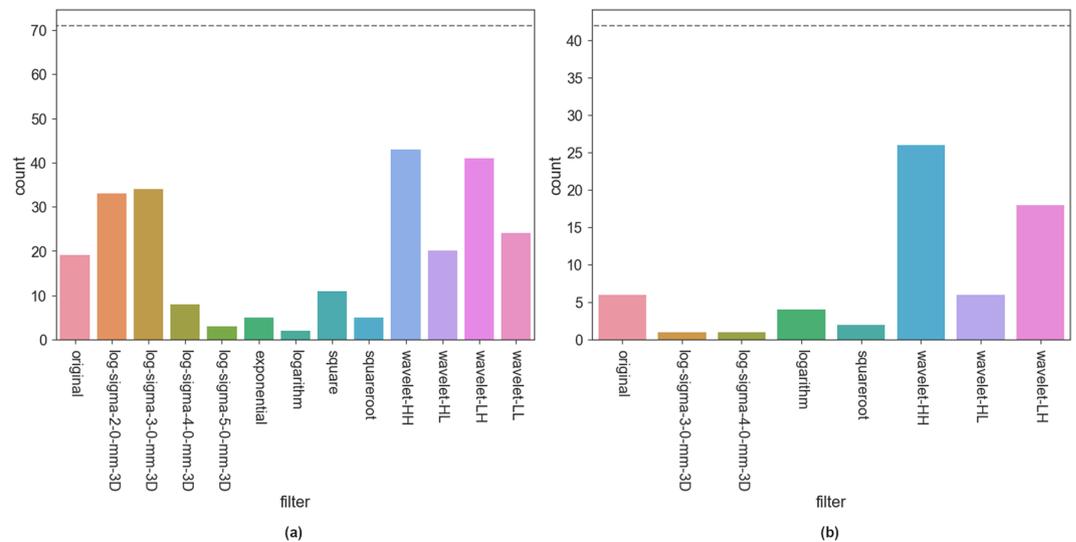
**Figure 6.** Top 3 features for each feature group by ICC in ADC images for (a) the Tumor ROI and (b) the Peripheral Zone. Results in these figures show again a wide spread of ICCs in the Tumor ROI, with some filters having tendency towards high ICCs, however, not consistently. For the Peripheral Zone many filters reach consistently above reference repeatability. Overall, ICCs are also much less spread. See Supplementary Dataset 1 for confidence intervals for the ICCs in this figure.

invariant under any of the investigated pre-processing options. Furthermore, these shape features are correlated to *Volume* and capture less information about shape than *Volume*. Hence most of these features are not likely to add any information that is not captured by *Volume* already.

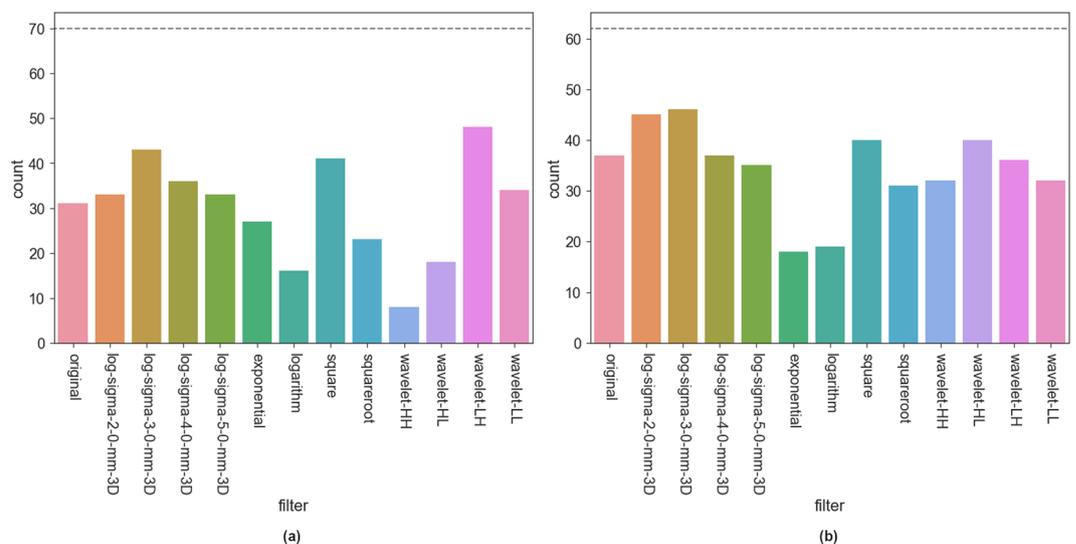
Even though we could not find configurations which consistently improved the repeatability for all or most features, we still found many features, which - under certain configurations - have a better repeatability than our reference. For the Tumor ROI about 70 features have a higher ICC than *Volume* on T2w as well as on ADC (see Figs 7 and 8). This could indicate that different features simply require different configurations. However, we don't believe this is the case, considering the small study size plus the fact that no obvious pattern emerges between certain pre-processing configurations and sub-groups of features. Furthermore, we were not able to fully reproduce the good repeatability of features reported in literature for PCA mpMRI on our dataset. This is another indication that there are many factors - even beyond the ones we assessed in this study - that influence the repeatability of features.

There are multiple factors that affect feature repeatability. An important one is the consistency in identification of the region of interest between the two timepoints. Even though the ROIs used in this study were contoured by an expert radiologist with 15+ years of experience in prostate MRI, we can expect that the segmentations will not be perfectly consistent between time points. Figure 1 illustrates that it is not always clear if the segmented regions really completely represent the same tissue region. However, a perfect match of manually defined ROIs cannot be expected in practice. The potential inconsistencies in region definition reflect the challenges of utilizing radiomics for decision support that might be encountered in clinical practice.

In our study, we made the decision to use ICC as the measure for evaluating repeatability. We use the ICC because it is commonly used in radiomics studies, and as such is a *de facto* convention in the field (e.g., see<sup>3,23,24</sup>). It also has the advantages of being invariant with respect to linear scaling and shifting. Some authors indicate



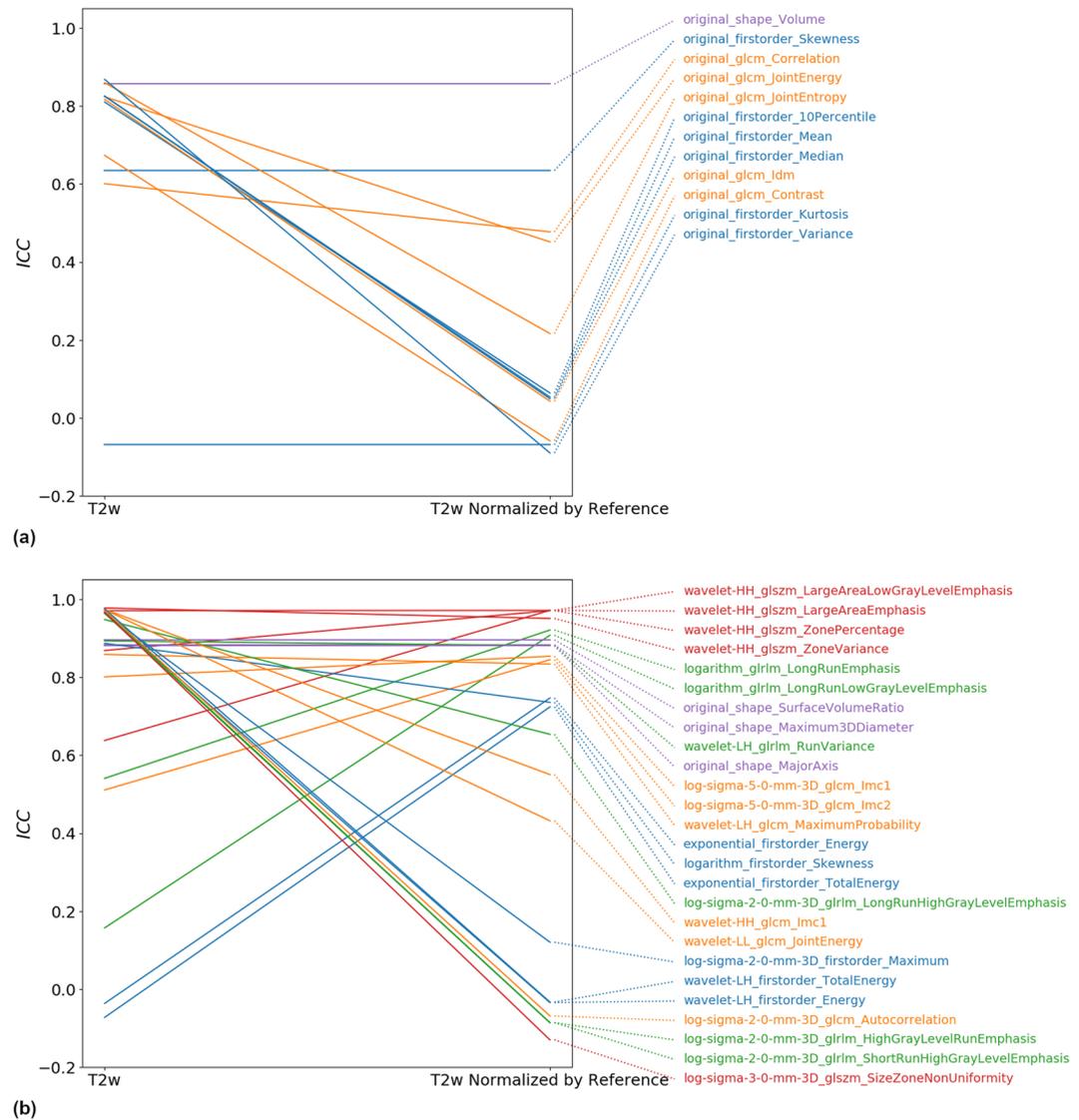
**Figure 7.** Overview of how often the particular pre-filters appear among the features which reach an ICC higher than Volume on T2w images for (a) Tumor ROI and (b) Peripheral Zone. Dashed line indicates total number of features with an ICC higher than *Volume*. Note that for one feature several filters can appear. Results in these figures illustrate that some filters are consistently more often related to high repeatability than others. However, no filter comes even close to being always related to high repeatability.



**Figure 8.** Overview of how often the particular pre-filters appear among the features which reach an ICC higher than Volume on ADC images for (a) Tumor ROI and (b) Peripheral Zone. Dashed line indicates total number of features with an ICC higher than *Volume*. Note that for one feature several filters can appear. Results in these figures illustrate that almost all filters are related to high repeatability in about half of the above-reference group. However, no filter comes close to being always related to high repeatability.

fixed thresholds for interpreting the ICC<sup>23,34</sup>. However, these can only be valid as a reference if our population variance was expected to be comparable to the ones used in other studies. In general, this is not the case though. As Raunig *et al.* note<sup>51</sup>: “ICC values for a very heterogeneous subject sample may yield very [sic] nearly perfect correlation based solely on the between-subject variance”. Hence, ICCs of different radiomics features over the same population of subjects should be compared to a reference within this population. Considering those limitations of the absolute ICC threshold, we use the *Volume* ICC as the reference. Tumor volume is an important quantitative measure characterizing PCa, which has been investigated earlier<sup>16</sup>, and evaluation of its repeatability in this specific dataset has already been presented in our earlier work<sup>42</sup>.

Another measure which is commonly used to assess repeatability is the repeatability coefficient (RC)<sup>51,59</sup>. However, the RC is not invariant with respect to scaling and it is denoted in absolute units of the feature. If two features are not expressing their values in the same units and scale space, it is not valid to compare them based on



**Figure 9.** Change in Tumor ROI ICCs from T2w to T2w normalized by a muscle reference region for (a) literature recommended features, (b) 3 most stable features from each of the feature classes. Note that for each image configuration in (b) the top 3 features were selected (hence up to 6 features per feature group are plotted). Results in this figure show that for many features, which have a high repeatability without normalization, normalization by a muscle reference region decreases repeatability. For some features, however, muscle-reference normalization improves repeatability.

their RC. Since most radiomics features are abstract measures with no direct real-world interpretation we cannot assume that they operate in the same unit and scales space and thus cannot use the RC for comparing different features. The RC is rather designated for assessing the agreement of two methods, measuring the same quantity<sup>59</sup>, and exposing the expected absolute differences between repeated measurements (limits of agreement).

Our study has limitations. A number of them are inherent to the dataset we used to extract the features (the limitations of the study that generated the MRI data and annotations are discussed extensively by Fedorov *et al.* in<sup>42</sup>): small sample size, small volumes of the identified tumors, lack of the analysis of multi-reader consistency. Specific to our analysis, we did not consider rigorous statistical modeling and testing for our evaluation in this study for several reasons. First, the sample size is rather small. Nevertheless, since there are no existing studies investigating repeatability of the mpMRI features in the prostate, we argue our results are nevertheless of value for the radiomics community. Second, the intention of this paper is an overview of the effects of pre-processing variations on the repeatability of radiomics features. A thorough statistical analysis of all the variations considered in this paper would extend the scope. Since all data of this study is available, we hope this will encourage researchers to perform rigorous statistical analysis as an extension to this study.

Evaluation of diagnostic performance of the features is out of scope of this manuscript. This is due to the small sample size, large number of features, non-binary nature of characterization of the disease by the radiologist (suspicion characterization was done using the 5-point PI-RADS v2 scale, as discussed in<sup>42</sup>), and lack of targeted

biopsy samples allowing to confirm the cancer grade for the annotated tumor regions. We also note that evaluation of repeatability is important on its own, as discussed in<sup>60</sup>. As an example, lung CT RIDER dataset was used to study repeatability of image-derived measures, and did not consider diagnostic performance, see studies by Zhao *et al.*<sup>61,62</sup>. Understanding of the repeatability of the features derived from that dataset has been instrumental for feature selection in numerous radiomics studies, such as<sup>3,63</sup>.

Inherently, our conclusions are specific to the dataset, image preprocessing and feature extraction approach, and the ICC as a measure of repeatability. There is an extensive variety of widely used repeatability measures in statistics, many of which are comprehensively summarized by Barnhart *et al.*<sup>60</sup>, and others proposed by imaging researchers, such as those used by Chirra *et al.*<sup>64</sup>. Our goal, however, was not to perform a comprehensive comparison of repeatability measures, but to keep the scope of this study focused, and utilize a measure that is already broadly accepted in the radiomics community. Similarly, it was not our goal to comprehensively explore the various options for image intensity normalization. Although various approaches to intensity normalization have been proposed (e.g., see Nyul *et al.*<sup>65</sup>), they may be applicable under assumptions that do not hold for our data (e.g., absence of pathology or imaging artifacts<sup>66</sup>). Perhaps more importantly, we stress that the normalization approaches we utilize (bias correction, whole-image normalization, and muscle-reference normalization) are employed in prominent recent studies in prostate MRI radiomics (see Bonekamp *et al.*<sup>20</sup> and Fehr *et al.*<sup>4</sup>). Investigation of alternative approaches to normalization and agreement assessment are certainly justified, but were deemed outside of the scope of the present study.

## Conclusion

Our study shows that radiomics features, as evaluated on the specific PCa mpMRI dataset we used, vary greatly in their repeatability. Furthermore, repeatability of radiomics features evaluated using ICC is highly susceptible to the processing configuration. Even on our small study population, the results already show that the type of image, preprocessing, and region of interest used to evaluate the feature can vastly change the repeatability of certain features. This could contribute to the explanation of why feature recommendations among recent studies are not consistent, and why we could not confirm good repeatability for some of the literature reported features.

We suggest caution when utilizing prior studies as a basis for pre-selection of radiomics features to be used in radiomics signatures. Whenever possible, repeatability analysis on a representative dataset should be done as part of the study-specific feature selection procedure. The dataset used in our study is publicly available<sup>40</sup>, and can be used to facilitate such feature selection procedures for the study-specific radiomics feature extraction procedures. If repeatability analysis is not possible or is impractical, and prior evidence is used for feature pre-selection, we recommend paying close attention to the reported configuration of the feature extraction process. Furthermore, it is important to consider whether the relevance of the assumptions on the reported study population are also valid for the planned study.

When publishing findings on radiomics studies (be it on repeatability or any other performance measure) we strongly advocate reporting of all the details describing the preprocessing and feature extraction procedures. To increase reproducibility of study findings we also strongly recommend following the consensus definitions of features (such as those proposed by the IBSI initiative)<sup>29</sup> and making the implementation available. When sharing of the dataset is not possible, we recommend that the pertinent details about the study population are reported to help with the interpretation of conclusions of the study. As one example, it is common that distribution of the tumor volumes is not summarized, although it is widely recognized that the size of the region of interest has a strong effect on the measurements extracted. To support reusability and further investigation, our study utilized a publicly available dataset<sup>40</sup>, and the open source radiomics library *pyradiomics*<sup>41</sup>. The calculated radiomics features and evaluation scripts are also published as open-source. This makes it possible to apply alternative implementations of radiomics feature extraction tools to the same dataset, and compare the result with the radiomics features evaluated in this study.

For the specific dataset and radiomics feature extraction approach considered in this study, we were not able to determine a general set of universally stable feature and preprocessing recommendations. Nevertheless, we found many features with a considerably higher repeatability than our reference (*Volume*). Most prominently these include the 3 top performing features for each feature group. Based on the analysis of our data, these can be strong candidates for inclusion into radiomics signatures. We consider a specific study of the predictive power of these candidates on a different dataset valuable future work.

## Data Availability

The MRI datasets analyzed in this study are from the QIN-PROSTATE-Repeatability TCIA collection<sup>40</sup>.

The extracted features used for all analyses in this paper are located at: <https://github.com/QIICR/QIN-PROSTATE-Repeatability-Radiomics/tree/master/EvalData>.

## Code Availability

*Pyradiomics* is available at: <https://github.com/Radiomics/pyradiomics> (the specific version used in this study is v1.3.0.post59 + g2e6d2c1, available at <https://github.com/Radiomics/pyradiomics/tree/2e6d2c1>).

Code to perform the analysis and figure generation is available at: <https://github.com/QIICR/QIN-PROSTATE-Repeatability-Radiomics>.

This repository also contains example code on how the features were extracted from the image data.

## References

1. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248 (2012).
2. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
3. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).

4. Fehr, D. *et al.* Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proc. Natl. Acad. Sci. USA* **112**, E6265–73 (2015).
5. Wibmer, A. *et al.* Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *Eur. Radiol.*, <https://doi.org/10.1007/s00330-015-3701-8> (2015).
6. Peng, Y. *et al.* Quantitative Analysis of Multiparametric Prostate MR Images: Differentiation between Prostate Cancer and Normal Tissue and Correlation with Gleason Score—A Computer-aided Diagnosis Development Study. *Radiology*, <https://doi.org/10.1148/radiol.13121454> (2013).
7. Cameron, A., Khalvati, F., Haider, M. A. & Wong, A. MAPS: A Quantitative Radiomics Approach for Prostate Cancer Detection. *IEEE Trans. Biomed. Eng.* **63**, 1145–1156 (2016).
8. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, 13087 (2015).
9. Parmar, C. *et al.* Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Front. Oncol.* **5**, 272 (2015).
10. Jain, R. *et al.* Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology* **272**, 484–493 (2014).
11. Gevaert, O. *et al.* Glioblastoma Multiforme: Exploratory Radiogenomic Analysis by Using Quantitative Image Features. *Radiology* **276**, 313 (2015).
12. Carvalho, S. *et al.* Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. *Acta Oncol.* **52**, 1398–1404 (2013).
13. Alic, L., Niessen, W. J. & Veenland, J. F. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. *PLoS One* **9**, e110300 (2014).
14. Barentsz, J. O. *et al.* ESUR prostate MR guidelines 2012. *Eur. Radiol.* **22**, 746–757 (2012).
15. Weinreb, J. C. *et al.* PI-RADS Prostate Imaging – Reporting and Data System: 2015, Version 2. *Eur. Urol.* **69**, 16–40 (2016).
16. Turkbey, B. *et al.* Correlation of magnetic resonance imaging tumor volume with histopathology. *J. Urol.* **188**, 1157–1163 (2012).
17. Hambrock, T. *et al.* Relationship between apparent diffusion coefficients at 3.0-T MR imaging and Gleason grade in peripheral zone prostate cancer. *Radiology* **259**, 453–461 (2011).
18. Langer, D. L. *et al.* Prostate Tissue Composition and MR Measurements: Investigating the relationship between ADC, T2, Ktrans, Ve and corresponding histological features. *Radiology* **255**, 485–494 (2010).
19. Smith, C. P. *et al.* Radiomics and radiogenomics of prostate cancer. *Abdominal Radiology*, <https://doi.org/10.1007/s00261-018-1660-7> (2018).
20. Bonekamp, D. *et al.* Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values. *Radiology* 173064, <https://doi.org/10.1148/radiol.2018173064> (2018).
21. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* **6**, 23428 (2016).
22. Balagurunathan, Y. *et al.* Test-retest reproducibility analysis of lung CT image features. *J. Digit. Imaging* **27**, 805–823 (2014).
23. Leijenaar, R. T. H. *et al.* Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol.* **52**, 1391–1397 (2013).
24. Gudmundsson, S., Runarsson, T. P. & Sigurdsson, S. Test-retest reliability and feature selection in physiological time series classification. *Comput. Methods Programs Biomed.* **105**, 50–60 (2012).
25. O'Connor, J. P. B. *et al.* Imaging biomarker roadmap for cancer studies. *Nat. Rev. Clin. Oncol.*, <https://doi.org/10.1038/nrclinonc.2016.162> (2016).
26. Parmar, C. *et al.* Robust Radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One* **9**, e102107 (2014).
27. Hunter, L. A. *et al.* High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. *Med. Phys.* **40** (2013).
28. Bologna, M. *et al.* Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images. *J. Digit. Imaging*, <https://doi.org/10.1007/s10278-018-0092-9> (2018).
29. Zwanenburg, A., Leger, S., Vallières, M., Löck, S. & for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv [cs.CV]*. at <http://arxiv.org/abs/1612.07003> (2016).
30. Zhao, B. *et al.* Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* **252**, 263–272 (2009).
31. Sadinski, M. *et al.* Short-term reproducibility of apparent diffusion coefficient estimated from diffusion-weighted MRI of the prostate. *Abdom. Imaging* **40**, 2523–2528 (2015).
32. Alonzi, R. *et al.* Reproducibility and correlation between quantitative and semiquantitative dynamic and intrinsic susceptibility-weighted MRI parameters in the benign and malignant human prostate. *J. Magn. Reson. Imaging* **32**, 155–164 (2010).
33. Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 1143–1158 (2018).
34. Hu, P. *et al.* Reproducibility with repeat CT in radiomics study for rectal cancer. *Oncotarget* **7**, 71440–71446 (2016).
35. van Velden, F. H. P. *et al.* Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol. Imaging Biol.* **18**, 788–795 (2016).
36. 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.* **58**, 406–411 (2017).
37. Emaminejad, N. *et al.* The effects of variations in parameters and algorithm choices on calculated radiomics feature values: initial investigations and comparisons to feature variability across CT image acquisition conditions. In *SPIE Medical Imaging 2018: Computer-Aided Diagnosis*, <https://doi.org/10.1117/12.2293864> (eds Mori, K. & Petrick, N.) 140 (SPIE, 2018).
38. Chalkidou, A., O'Doherty, M. J. & Marsden, P. K. False discovery rates in PET and CT studies with texture features: A systematic review. *PLoS One* **10**, 1–18 (2015).
39. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Phys. Med. Biol.* **61**, R150–R166 (2016).
40. Fedorov, A. *et al.* An annotated test-retest collection of prostate multiparametric MRI. *Scientific Data* **5**, 180281 (2018).
41. van Griethuysen, J. J. M. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **77**, 104–108 (2017).
42. Fedorov, A., Vangel, M. G., Tempany, C. M. & Fennessy, F. M. Multiparametric Magnetic Resonance Imaging of the Prostate: Repeatability of Volume and Apparent Diffusion Coefficient Quantification. *Invest. Radiol.* **52**, 538–546 (2017).
43. Somford, D. M. *et al.* Initial Experience With Identifying High-Grade Prostate Cancer Using Diffusion-Weighted MR Imaging (DWI) in Patients With a Gleason Score  $\leq 3 + 3 = 6$  Upon Schematic TRUS-Guided Biopsy: A Radical Prostatectomy Correlated Series. *Invest. Radiol.* **47**, 153 (2012).
44. Fedorov, A. *et al.* 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).
45. Rios Velazquez, E. *et al.* Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer. *Cancer Res.* **77**, 3922–3930 (2017).
46. Tustison, N. J. *et al.* N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).

47. Huang, W. *et al.* The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge. *Tomography* **2**, 56–66 (2016).
48. Leijenaar, R. T. H. *et al.* The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Sci. Rep.* **5**, 11075 (2015).
49. Tixier, F. *et al.* Reproducibility of Tumor Uptake Heterogeneity Characterization Through Textural Feature Analysis in 18F-FDG PET. *J. Nucl. Med.* **53**, 693–700 (2012).
50. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428 (1979).
51. Raunig, D. L. *et al.* Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment. *Stat. Methods Med. Res.* **24**, 27–67 (2015).
52. Schwier, M. *et al.* Repeatability of Multiparametric Prostate MRI Radiomics Features. *arXiv [cs.CV]* at, <http://arxiv.org/abs/1807.06089> (2018).
53. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **27**, 832–837 (1956).
54. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33**, 1065–1076 (1962).
55. Kuhl, C. K. *et al.* Abbreviated Biparametric Prostate MR Imaging in Men with Elevated Prostate-specific Antigen. *Radiology* 170129, <https://doi.org/10.1148/radiol.2017170129> (2017).
56. Malyarenko, D. *et al.* Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. *J. Magn. Reson. Imaging* **37**, 1238–1246 (2013).
57. Kwon, D. *et al.* Classification of suspicious lesions on prostate multiparametric MRI using machine learning. *J. Med. Imag.* **5**, 034520 (2018).
58. Wang, H. J. *et al.* Value of normalized apparent diffusion coefficient for estimating histological grade of vesical urothelial carcinoma. *Clin. Radiol.* **69**, 727–731 (2014).
59. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310 (1986).
60. Barnhart, H. X. *et al.* Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. *Stat. Methods Med. Res.* 0962280214534651–, <https://doi.org/10.1177/0962280214534651> (2014).
61. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* **6**, 1–7 (2016).
62. Zhao, B. *et al.* Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non-Small Cell Lung Cancer. *Radiology* **252**, 263–272 (2009).
63. Vaidya, P. *et al.* RaPtomics: integrating radiomic and pathomic features for predicting recurrence in early stage lung cancer. In *Medical Imaging 2018: Digital Pathology* **10581**, 105810M (International Society for Optics and Photonics, 2018).
64. Chirra, P. *et al.* Empirical evaluation of cross-site reproducibility in radiomic features for characterizing prostate MRI. In *Medical Imaging 2018: Computer-Aided Diagnosis* **10575**, 105750B (International Society for Optics and Photonics, 2018).
65. Nyúl, L. G., Udupa, J. K. & Zhang, X. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging* **19**, 143–150 (2000).
66. Shinohara, R. T. *et al.* Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing & Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin* **6**, 9–19 (2014).

## Acknowledgements

Funding support: NIH U24 CA180918, U01 CA151261, P41 EB015898, U24 CA194354, and U01 CA190234.

## Author Contributions

A.F. and M.S. conceived and conducted of the project, analysed the data, and wrote the paper. J.V.G. provided expert guidance on radiomics feature extraction and data analysis, and reviewed the manuscript. M.G.V., S.Pi., S.Pe., C.T., H.J.W.L.A., R.K. and F.M.F. provided expert guidance on data analysis and medical questions, and reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-019-45766-z>.

**Competing Interests:** Clare Tempny declares ongoing financial relationship with Profound Medical. All other authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019