cancers

Article

Image Contrast, Image Pre-Processing, and T_1 Mapping Affect MRI Radiomic Feature Repeatability in Patients with Colorectal Cancer Liver Metastases

Damien J. McHugh [1,2], Nuria Porta [3], Ross A. Little [1,2], Susan Cheung [1,2], Yvonne Watson [1,2], Geoff J. M. Parker [4,5], Gordon C. Jayson [1,6] and James P. B. O'Connor [1,2,7,8,*]

[1] Division of Cancer Sciences, The University of Manchester, Manchester M13 9PL, UK; damien.mchugh@manchester.ac.uk (D.J.M.); ross.little@manchester.ac.uk (R.A.L.); susan.cheung@manchester.ac.uk (S.C.); yvonwatson2@gmail.com (Y.W.); gordon.jayson@manchester.ac.uk (G.C.J.)
[2] Quantitative Biomedical Imaging Laboratory, The University of Manchester, Manchester M13 9PL, UK
[3] Clinical Trials and Statistics Unit, Institute of Cancer Research, London SW3 6JB, UK; nuria.porta@icr.ac.uk
[4] Centre for Medical Image Computing, University College London, London WC1V 6LJ, UK; geoff.parker@ucl.ac.uk
[5] Bioxydyn Ltd., Manchester M15 6SZ, UK
[6] Department of Medical Oncology, The Christie Hospital, Manchester M20 4BX, UK
[7] Department of Radiology, The Christie Hospital, Manchester M20 4BX, UK
[8] Division of Radiotherapy and Imaging, Institute of Cancer Research, London SW3 6JB, UK
* Correspondence: james.oconnor@icr.ac.uk

Simple Summary: Medical images are data. They contain more information than is routinely identified by radiologists reading scans. Many scientists are investigating if extracting shape and grey-scale features from images can predict which oncology patients will respond to therapy. This approach, termed 'radiomics', must be validated before being ready for clinical use. One step is to determine measurement repeatability to ensure that radiomic features are robust, and that changes in features reflect genuine changes in tumours. In this study patients had two repeated sets of magnetic resonance imaging scans. We found that radiomic feature repeatability varied depending on scan acquisition parameters and the use of an administered contrast agent. We also compared how different repeatability assessment methods can best reveal these findings. We conclude that measuring radiomic feature repeatability is essential, but is also complex and prone to pitfalls. Overall, our study provides several insights into how radiomic feature repeatability is best assessed.

Abstract: Imaging biomarkers require technical, biological, and clinical validation to be translated into robust tools in research or clinical settings. This study contributes to the technical validation of radiomic features from magnetic resonance imaging (MRI) by evaluating the repeatability of features from four MR sequences: pre-contrast T_1- and T_2-weighted images, pre-contrast quantitative T_1 maps (qT_1), and contrast-enhanced T_1-weighted images. Fifty-one patients with colorectal cancer liver metastases were scanned twice, up to 7 days apart. Repeatability was quantified using the intraclass correlation coefficient (ICC) and repeatability coefficient (RC), and the impact of non-Gaussian feature distributions and image normalisation was evaluated. Most radiomic features had non-Gaussian distributions, but Box–Cox transformations enabled ICCs and RCs to be calculated appropriately for an average of 97% of features across sequences. ICCs ranged from 0.30 to 0.99, with volume and other shape features tending to be most repeatable; volume ICC $>$ 0.98 for all sequences. 19% of features from non-normalised images exhibited significantly different ICCs in pair-wise sequence comparisons. Normalisation tended to increase ICCs for pre-contrast T_1- and T_2-weighted images, and decrease ICCs for qT_1 maps. RCs tended to vary more between sequences than ICCs, showing that evaluations of feature performance depend on the chosen metric. This work suggests that feature-specific repeatability, from specific combinations of MR sequence and pre-processing steps, should be evaluated to select robust radiomic features as biomarkers in specific studies. In addition, as different repeatability metrics can provide different insights into a specific feature, consideration of the appropriate metric should be taken in a study-specific context.

Citation: McHugh, D.J.; Porta, N.; Little, R.A.; Cheung, S.; Watson, Y.; Parker, G.J.M.; Jayson, G.C.; O'Connor, J.P.B. Image Contrast, Image Pre-Processing, and T_1 Mapping Affect MRI Radiomic Feature Repeatability in Patients with Colorectal Cancer Liver Metastases. *Cancers* **2021**, *13*, 240. https://doi.org/10.3390/cancers13020240

Received: 7 December 2020
Accepted: 5 January 2021
Published: 11 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: radiomics; MRI; repeatability; repeatability coefficient; intraclass correlation coefficient; liver metastases

1. Introduction

Imaging underpins much of the current management of patients with cancer, through diagnosis, staging and monitoring response to therapy. There is considerable current interest in evaluating if high throughput analysis of medical images—in an approach termed 'radiomics'—can further extend the role of imaging by producing signatures that are prognostic or are predictive of clinical outcome [1–3].

For radiomics to yield robust imaging biomarkers, technical, biological, and clinical validation are required [4,5]. Such validation is a multi-step process, in which various aspects of a proposed biomarker's performance are evaluated. Technical validation requires the evaluation of biomarker accuracy, repeatability, reproducibility, and availability, while biological and clinical validation require an understanding of how features relate to underlying biology, and how they relate to outcome, respectively. There has been a substantial amount of research into how radiomic features relate to tumour biology and outcome. Many studies have focused on finding a statistical association between either one radiomic feature, or several features combined into a 'signature', and an underlying biological feature or clinical outcome [6,7]. For example, computed tomography (CT) radiomic features from the Gray-Level Run Length Matrix in the tumour and peripheral ring, along with the minimum value in the tumour, have been associated with CD8 cell infiltration across a range of tumour types [8], and CT features related to tumour heterogeneity and compactness/sphericity have shown an association with survival in lung and head and neck cancer [2].

In terms of technical validation, measuring repeatability in single centres and reproducibility across multiple centres is crucial, and provides an important step in developing metrology standards for quantitative imaging biomarkers in general [9], including radiomic features [10]. While a number of studies have assessed the repeatability and/or reproducibility of CT-derived cancer radiomic features [11], there are fewer studies investigating the repeatability of MR-derived radiomic features [12–21]. These studies are limited by small patient numbers ($\leq$17, with the exception of Kickingereder et al., [13] and Merisaari et al. [20], with 55 and 112 patients, respectively), and the use of only one or two MR sequences (except in [13] where three were used). Furthermore, given the difficulty in directly comparing MR signal intensities from different scans, the effect of image normalisation on repeatability needs to be considered for different MR sequences, as does the validity of assumptions underlying the statistical analysis of repeatability. Finally, the impact of gadolinium-based contrast agents on the repeatability of features from T_1-weighted images, and the repeatability of features from T_1 maps, is yet to be evaluated in tumours.

This study aimed to address these knowledge gaps. To achieve this we sought to provide a comprehensive evaluation of the repeatability of MRI derived radiomic features in patients with liver metastases from colorectal cancer, which is an emerging clinical site of interest [22].

2. Materials and Methods

2.1. Image Acquisition and ROI Definition

We examined pre-treatment MRI scan data in patients recruited for a clinical trial (EudraCT number 2009-011377-33) [23]. All patients gave written informed consent. The study was conducted in accordance with the Declaration of Helsinki, and the study received institutional board approval (North West–Greater Manchester Central Research Ethics Committee, REC 09/H1008/99). All included patients were aged 18 years or over, had primary colorectal cancer, with at least one liver metastasis measuring >2.5 cm in maximum dimension on their screening CT scan, had a performance status of 0–2, and had two pre-treatment scans (median time between scans = 4 days, range = 2–7 days) (Figure 1a). Given the retrospective nature of this study, the available data determined the sample size, rather than a formal power calculation.

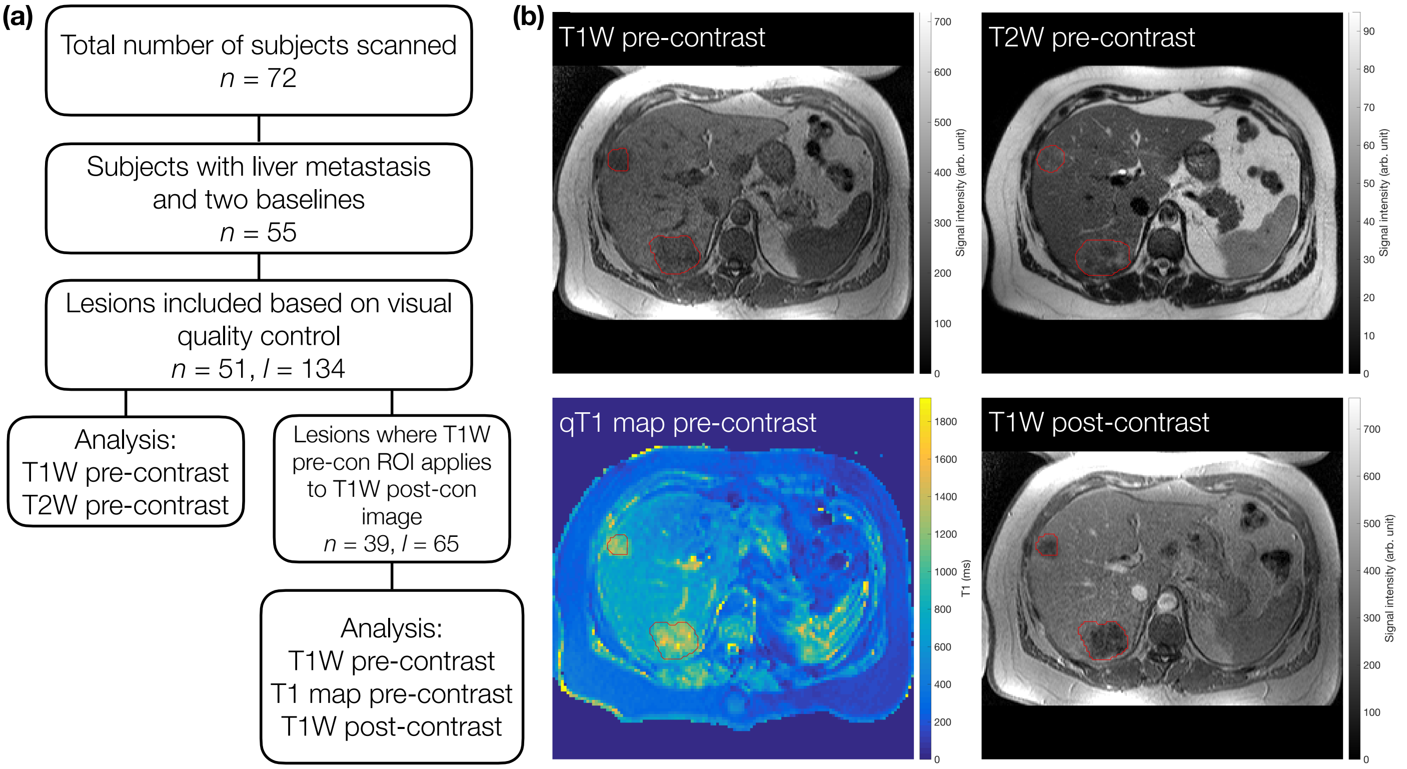

Figure 1. (**a**) Flowchart for subject and lesion selection. For T_1W and T_2W pre-contrast images, a total of 51 subjects (*n*) with 134 lesions (*l*) were included. Sixty-five of these lesions were also analysed on T_1W post-contrast images and pre-contrast quantitative T_1 maps. For direct comparison in terms of number of lesions included, the T_1W pre-contrast images from this subset were re-analysed in 39 patients. (**b**) Example images for one subject with two metastases (red outlines).

MRI acquisition and analysis was performed to Good Clinical Practice (GCP) standards. Data were acquired on a 1.5 T Philips Achieva scanner (Philips Healthcare, Best, The Netherlands). All imaging sequences were acquired without breath-holds or gating, with 25 axial slices, either 4 or 8 mm thick. Individual sequence parameters were:

- Multislice 2D T_1-weighted turbo field-echo sequence prior to contrast agent administration (flip angle (FA) = 15°, repetition time (TR) = 10 ms, echo time (TE) = 4.60 ms, field of view (FoV) = 375 mm × 264 mm, acquired in-plane resolution = 1.46 mm × 2.09 mm, reconstructed in-plane resolution = 1.46 mm × 1.46 mm, 25 slices); hereafter termed T_1W pre-contrast.
- Multislice 2D T_2-weighted turbo spin-echo sequence (FA = 90°, TR = 541 ms, TE = 80 ms, FoV = 375 mm × 264 mm, acquired in-plane resolution = 1.46 mm × 1.84 mm, reconstructed in-plane resolution = 1.46 mm × 1.46 mm, 25 slices); hereafter termed T_2W pre-contrast.
- Three 3D spoiled gradient-echo sequences (FAs = $2°, 10°, 20°$, TR = 4 ms, TE = 0.82 ms, FoV = 375 mm × 375 mm, acquired in-plane resolution = 2.93 mm × 3.71 mm, reconstructed in-plane resolution = 2.93 mm × 2.93 mm, 25 slices). T_1 was quantified by fitting the spoiled gradient-echo equation [24] to the signal intensities, S, in the three flip angle images: $S(\alpha) = M_0 \sin\alpha(1 - E)/(1 - E\cos\alpha)$, where M_0 is a factor proportional to proton density, α is the flip angle, and $E = \exp(-\mathrm{TR}/\mathrm{T}_1)$. Fitting was performed on a voxelwise basis using a Levenberg-Marquardt algorithm, and the resulting T_1 maps are hereafter termed qT_1 maps.

- Multislice 2D T_1-weighted turbo field-echo sequence acquired 5 min after contrast agent administration (FA = 15°, TR = 10 ms, TE = 4.60 ms, FoV 375 mm × 264 mm, acquired in-plane resolution = 1.46 mm × 2.09 mm, reconstructed in-plane resolution = 1.46 mm × 1.46 mm, 25 slices). 0.1 mmol/kg of gadoterate meglumine contrast agent (Dotarem, Guebert, France) was administered intravenously at a rate of 3 ml/s, using a Medrad Spectris power injector (Bayer, Reading, UK); hereafter termed T_1W post-contrast.

Regions of interest (ROIs) were defined manually using Java Image software (JIM version 6.0_16, Xinpase Systems Ltd, UK) by a radiologist (J.P.B.O'C; 16 years of experience). In each case an ROI was drawn on the T_1W pre-contrast images, and again on the T_2W pre-contrast images. Both pre-treatment scans were annotated together. Up to five target lesions were identified for each patient. Next, the qT_1 maps and the T_1W post-contrast images were inspected along with the ROIs drawn on the accompanying T_1W pre-contrast image to determine if the ROIs provided accurate delineation of each target lesion on these sequences. If the ROI was not deemed accurate for the qT_1 maps and the T_1W post-contrast images then data were excluded for those target lesions. Lesions included for the qT_1 maps and T_1W post-contrast images were therefore a subset of those for the T_1W and T_2W pre-contrast images; the T_1W pre-contrast images from this subset were treated as another dataset, to allow a direct comparison (in terms of lesion numbers) with qT_1 map and T_1W post-contrast data (Figure 1a). The qT_1 map masks were created based on the ROI defined on the higher-resolution T_1W pre-contrast images. Due to the different resolutions of the weighted and quantitative images, these qT_1 map masks were subsequently up-sampled for application to the subset of T_1W pre-contrast images, to ensure directly comparable masks were used when comparing qT_1 maps to the T_1W pre-contrast subset (Supplementary Information Figure S1). Each patient therefore had data for up to four sequences for two separate scan sessions. Example ROIs for one patient are shown in Figure 1b.

2.2. Radiomic Feature Extraction

For all ROI in all imaging sequences, radiomic features were extracted using PyRadiomics, version 2.2.0.post41+gc46ed88 [25]. PyRadiomics was chosen as it is open source, which facilitates reproducible research, and is largely ISBI-compliant (for details of differences see https://pyradiomics.readthedocs.io/en/latest/faq.html), which can aid comparison with studies using other ISBI-compliant platforms [26]. In total, 105 features were extracted from every lesion, with each feature belonging to one of the seven standard classes in PyRadiomics: Shape, First Order, Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), and Neighbouring Gray Tone Difference Matrix (NGTDM). Descriptions and formulae for all features can be found at https://pyradiomics.readthedocs.io/en/latest/features.html. 3D feature extraction was performed for images without normalisation, and for images normalised to have a mean of 0 and standard deviation of 100, using PyRadiomics' standard linear normalisation. In all cases, a bin width of 5 was used, and no filters were applied to images before feature extraction. For T_1W pre-contrast, T_2W pre-contrast, qT_1 pre-contrast, and T_1W post-contrast, the median number of bins over all lesions and visits was 60, 13, 310, and 110 (non-normalised) and 29, 42, 45, and 38 (normalised). YAML parameter files and analysis code used in this study can be found at https://gitlab.com/manchester_qbi/manchester_qbi_public/radiomics_repeatability.

2.3. Repeatability Analysis and Statistical Comparison

The repeatability of radiomic features was evaluated using the intraclass correlation coefficient (ICC) [27–29] and the repeatability coefficient (RC) [30]. ICC is a measure of within-subjects consistency relative to the total variability observed in the population, and it can be estimated in different ways, depending on the underlying statistical model that best captures the data structure under study; indeed, a feature's ICC depends on the variability of the feature across the cohort studied [28]. Here, ICC(1,1) [28] was judged to be the most suitable (see Supplementary Information), with point estimates and 95% CIs calculated as described by McGraw and Wong [27]. RC point estimates and 95% CIs were calculated as described by Barnhart and Barboriak [30]. RC is proportional to the within-subject standard deviation, and for a given feature the difference between repeat measurements is expected to fall within $-$RC and $+$RC, for 95% of patients. As such, RC is a useful metric for determining significant changes in a feature over time, for example in assessing treatment-induced changes relative to baseline. In contrast to ICCs, RCs depend on the magnitude and unit of the underlying feature, and themselves have the unit that the feature is measured in.

The models underlying ICC and RC calculations assume the feature follows a Gaussian distribution [29,31], and deviations from this assumption have been reported to impact ICC point estimates and confidence intervals (CIs) [32]. The assumption of Gaussian feature distributions was tested using the Shapiro–Wilk test for all radiomic features extracted from all MR sequences (for non-normalised and normalised images). Feature distributions were judged to be non-Gaussian based on a Bonferroni-corrected $p < 0.05/105$ threshold in the Shapiro–Wilk test (given 105 features per dataset). For non-Gaussian features, the optimal λ parameter to use in a Box–Cox transformation was found, such that the original feature distribution, x, could be transformed to a new distribution, y, which was consistent with a Gaussian distribution, according to $y = (x^{\lambda} - 1)/\lambda$, if $\lambda \neq 0$; $y = \log(x)$, if $\lambda = 0$ [33]. Shapiro–Wilk tests and Box–Cox transformations were carried out using the *stats* subpackage of the SciPy library in Python [34].

Note that we refer to 'Gaussian distributions' of the radiomics features rather than 'normal distributions', to avoid confusion with the term 'image normalisation', which refers to the transformation applied to image signal intensities prior to feature extraction. To assess the impact of Box–Cox transformations and image normalisation across images with different magnitudes and units, only ICCs were used, as RCs depend on the magnitude and unit of the underlying feature. As such, there were 16 datasets from which ICCs could be obtained: 4 MR sequences, each without and with image normalisation, and each without and with applying Box–Cox transformations to feature distributions.

To compare repeatability across different sequences, both ICC and RC were estimated, but for RC, only normalised datasets were used (where signal intensities become dimensionless) and were only estimated on the subset of features where the application of the Box–Cox transformation was consistent for all sequences (that is, for features where the transformation either was applied for all sequences, or was not applied for any sequence, to ensure comparable feature magnitudes). It should be noted that for features where the Box–Cox transformation was used, the direct comparison of RCs across sequences is hampered due to the different optimal λ values for the different sequences, as λ influences the feature magnitudes. This may potentially confound interpretability when comparing RCs from Box–Cox transformed features.

ICCs were formally compared using Fisher's Z-test [35], which involves applying a transformation to the ICCs; note that this is independent of the Box–Cox transformations described above. The test allowed ICCs to be compared between features without and with Box–Cox transformations (e.g., T_1W pre-contrast without Box–Cox transformations vs. T_1W pre-contrast with Box–Cox transformations), between features from different MR sequences (e.g., T_1W pre-contrast vs. T_2W pre-contrast), and between features from non-normalised and normalised images (e.g., T_1W pre-contrast without normalisation vs. T_1W pre-contrast with normalisation). In these comparisons, ICCs were taken to be significantly different based on a Bonferroni-corrected $p < 0.05/105$ threshold. RCs were descriptively compared across sequences, and with ICCs, on a per-feature basis.

3. Results

3.1. Effect of Box–Cox Transformations

Figure 2 illustrates the Box–Cox transformation procedure for three example features from non-normalised T_1W pre-contrast images. The chosen features, GLCM Inverse

Difference Moment, GLSZM Large Area Low Gray Level Emphasis, and Elongation reflect three scenarios in terms of feature distributions: (a) the original distribution is Gaussian, and does not need transforming before ICC calculation; (b) the original distribution is not Gaussian, but becomes Gaussian after a Box–Cox transformation; and (c) the original distribution is not Gaussian, but is still not Gaussian after a Box–Cox transformation. In case (c), although the feature still fails the Shapiro–Wilk test, the quantile-quantile (Q-Q) plot indicates that the transformation does help to make the distribution more consistent with a Gaussian. Across all sequences, with and without normalisation, the majority of features reflected scenario (b), where the transformation corrected a previously non-Gaussian distribution (Table 1). With this procedure, Gaussian distributions could be obtained for >93% of features for all datasets. For features where the applied transformation still did not yield a Gaussian distribution, qualitative assessment of Q-Q plots suggested some improvement, so these transformed features were still used. Supplementary Information Figures S2–S5 show feature histograms, Box–Cox transformations, and Q-Q plots for all features, for all sequences.

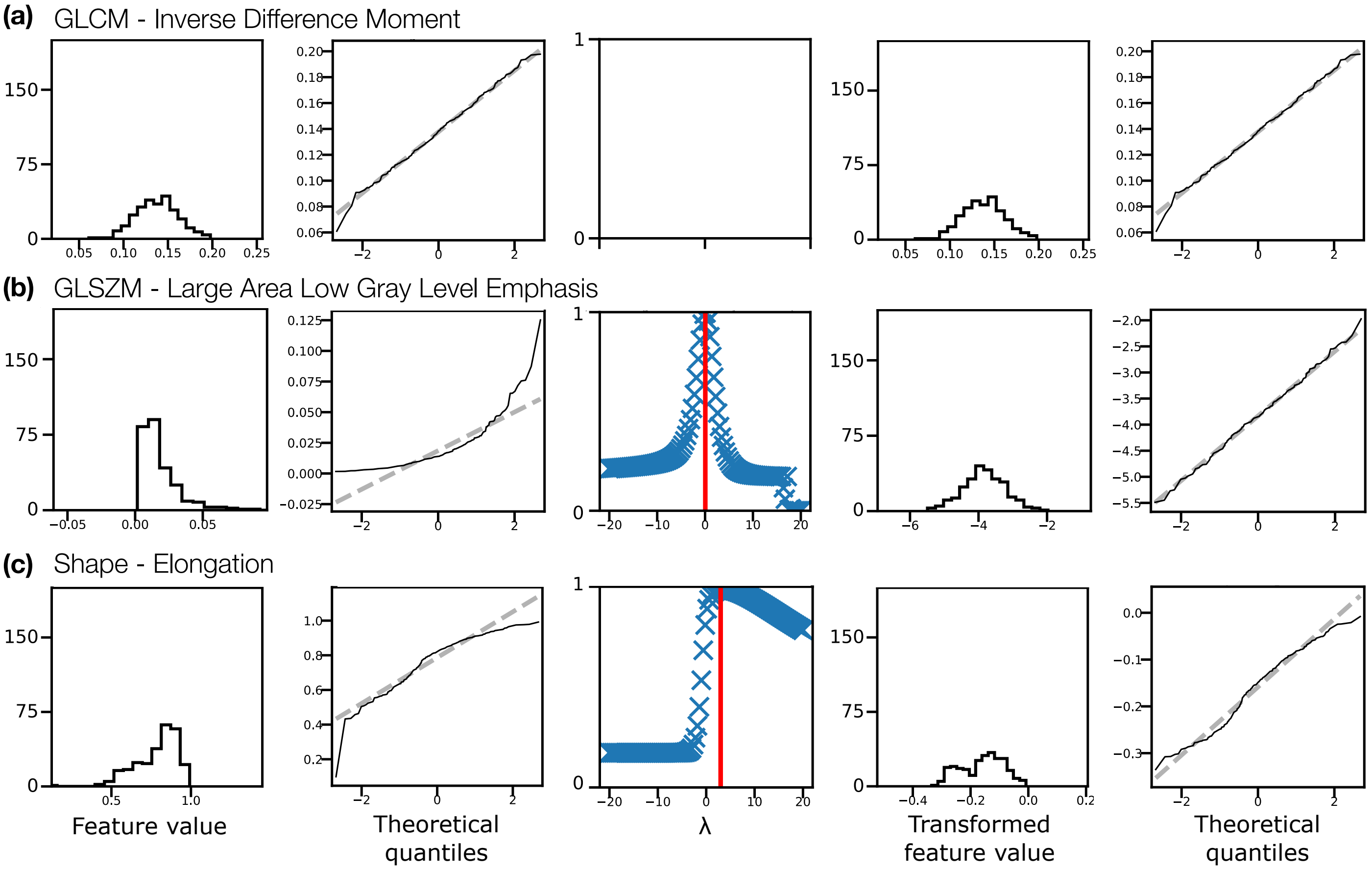

Figure 2. Feature distributions and transformations for three example features, (**a**) Inverse Difference Moment, (**b**) Large Area Low Gray Level Emphasis, (**c**) Elongation, from non-normalised T_1W pre-contrast images, pooling feature values from both visits for all subjects. Feature distributions and Q-Q plots for the original features are shown in the first and second columns, respectively. The third column shows the Box–Cox normality plots, with red vertical lines indicating the optimal λ to use to transform the distributions. Feature distributions and Q-Q plots for the transformed features are shown in the fourth and fifth columns, respectively. In (**a**) the original distribution is consistent with a Gaussian distribution (Shapiro–Wilk test), and does not require a transformation. In (**b**) the original distribution is not consistent with a Gaussian distribution, and the transformation corrects this. In (**c**) the original distribution is not consistent with a Gaussian distribution, but the transformed distribution is still not Gaussian, though the Q-Q plots suggest the transformed distribution is closer to Gaussian than the original. Y-axis labels are omitted for clarity, but are: Counts, Sample quantiles, Correlation coefficient, Counts, Sample quantiles, for each column, respectively.

Table 1. Effectiveness of Box–Cox transformations, assessed by the number of features with distributions consistent with a Gaussian distribution in different scenarios, for four MR sequences (rows), without and with normalisation (left and right). The first, second, and third column on each side show the number of features whose original distribution was Gaussian (Pre Box–Cox), the number of features whose original non-Gaussian distribution was transformed to Gaussian (Post Box–Cox), and the number of features whose distribution was non-Gaussian before and after applying transformations (Never). In all cases the total is 105, reflecting the number of features extracted.

	No Normalisation			**Normalisation**		
	Pre Box–Cox	**Post Box–Cox**	**Never**	**Pre Box–Cox**	**Post Box–Cox**	**Never**
T_1W pre-contrast	22	77	6	17	81	7
T_2W pre-contrast	17	84	4	15	87	3
qT_1 map pre-contrast	25	78	2	27	77	1
T_1W post-contrast	26	77	2	34	70	1

Figure 3 shows the effect of the Box–Cox transformations on calculated ICCs, for T_1W and T_2W pre-contrast non-normalised images (Supplementary Information Figure S6 shows equivalent plots for qT_1 maps and T_1W post-contrast non-normalised images). Note that these figures do not directly plot the ICC differences, but rather the difference in ICCs after applying Fisher's Z-transformation, and the error bars show the 95% CI on this difference. Note also that not all features where the 95% CIs do not contain zero are marked as significant due to the use of Bonferroni correction.

Depending on the feature, ICC point estimates could increase or decrease due to the transformation, though differences tended to be relatively small, and significant differences were observed for a minority of features. This suggests that while ICC calculations assume features follow a Gaussian distribution, there is a degree of robustness against cases where this assumption is invalid. While the application of the transformations does not have a dramatic effect on ICCs, it does help ensure the validity of methodological assumptions, and it can significantly affect the repeatability of some features. As such, ICCs from Box–Cox transformed features were used throughout the rest of the analysis.

3.2. ICC and RC Overview

Features exhibited a wide range of repeatabilities, with ICCs ranging from 0.30 (GLSZM Small Area Emphasis for T_2W images without normalisation) to 0.99 (Voxel Volume for T_1W post-contrast images). For all sequences, Voxel Volume and Mesh Volume were the features with the two highest ICCs ($>$0.98), indicating that tumour volume was the most repeatable feature.

Figure 4a provides an overview of the ICC results, plotting ICCs for all features, for each sequence, with and without normalisation, facilitating comparisons across all datasets. The dominance of yellow in the Shape class illustrates the tendency for higher ICCs for these features, though Sphericity has a notably lower ICC ($\sim$ 0.58) across all sequences. Excluding Shape features, nine features have ICC $>$ 0.90 for all sequences, with and without normalisation): Energy, Total Energy, GLRLM Gray Level Non Uniformity, GLRLM Run Length Non Uniformity, GLSZM Gray Level Non Uniformity, GLSZM Size Zone Non Uniformity, GLDM Dependence Non Uniformity, GLDM Gray Level Non Uniformity, and NGTDM Coarseness. However, all these features are correlated with Mesh Volume (absolute Spearman's ρ ranged from 0.62 to 0.97, across all sequences, with and without normalisation), suggesting these features offer limited independent information beyond tumour size. Figure 4b further summaries the impact of sequence choice and normalisation on repeatabilities, plotting the coefficient of variation (CoVs) in ICCs across all sequences, with and without normalisation. Shape features, with the exception of Sphericity, and NGTDM features tend to show the lowest variation, while features in other classes show greater variation in repeatability.

RC values for all features and datasets are shown in Supplementary Information Figures S7–S14. Note that in general these values cannot be directly compared for different features and sequences, due to their differing units. For the subset of Shape features where RC values can be compared, Figure 5 plots RC against ICC for the four sequences. RC tends to show more variability between sequences than ICC, with qT_1 and T_1W-post tending to have lower RCs. While ICCs tend to be uniformly high across sequences, indicating good repeatability regardless of sequence, the lower RCs for qT_1 map and T_1W-post suggest these would be preferred for detecting longitudinal changes in Shape features. ICCs and RCs can therefore be seen to offer different perspectives on repeatability, with the choice of metric informed by the context of a particular study. Supplementary Information Figures S15–S20 plot RC against ICC for the other feature classes. For some features, ICCs and RCs are inversely correlated, suggesting they provide comparable information about repeatability (with higher ICCs and lower RCs indicating better repeatability); for others, however, correlations are not observed, showing that for certain features, a low RC does not necessarily imply a high ICC. For example, Energy and Total Energy from qT_1 maps have relatively low RCs, but have the lowest ICCs of the four sequences. Conversely, for most other First Order features, qT_1 maps exhibit the highest RC and lowest ICC (Figure S15). Again, it should be noted that the direct comparison of RCs for Box–Cox transformed features is confounded by having different optimal λ values for the different sequences.

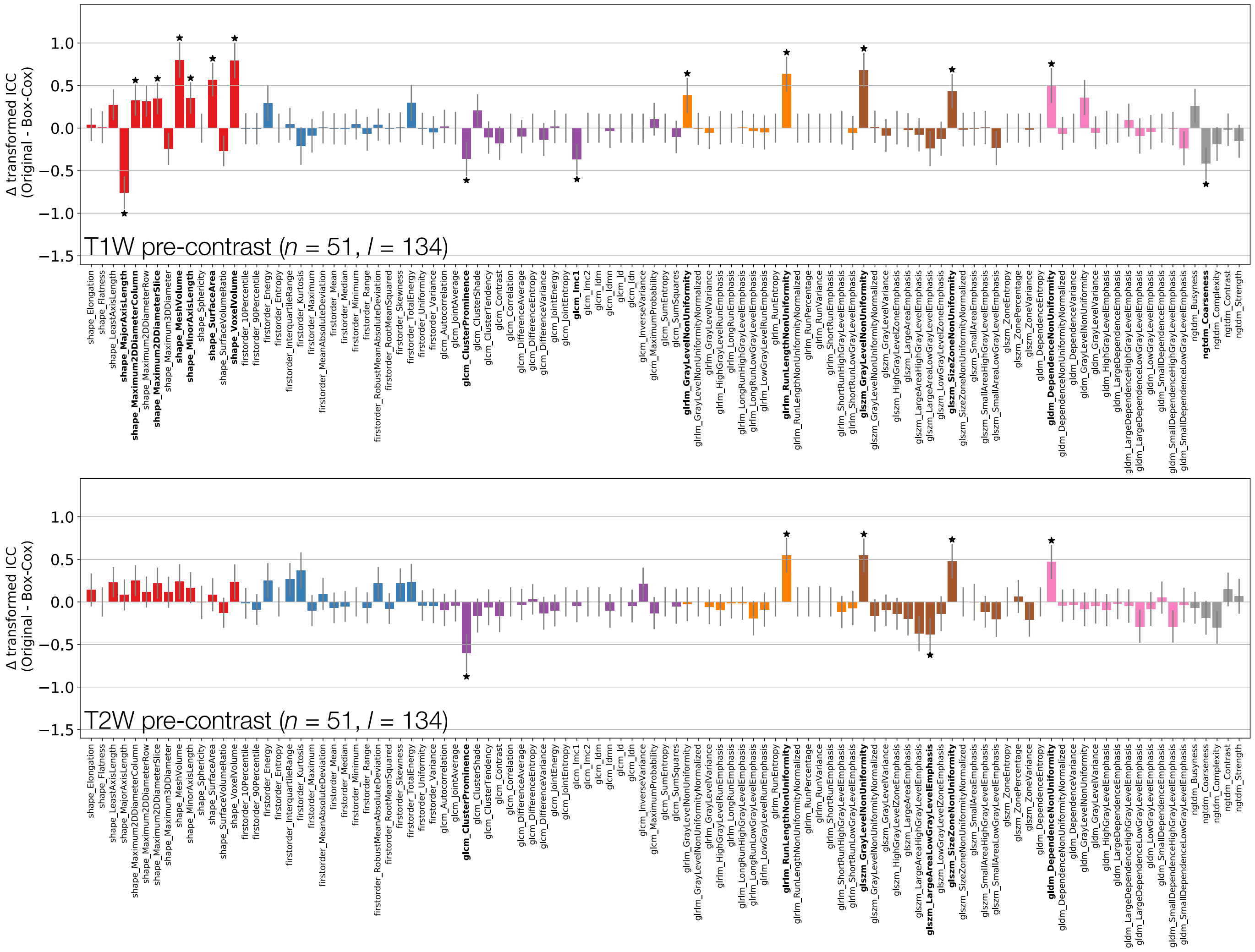

Figure 3. Effect of applying Box–Cox transformations to feature distributions, for T_1W (top) and T_2W (bottom) pre-contrast images. Bars represent the difference in ICC point estimates (after applying Fisher's Z-transformation), and error bars represent 95% CIs. Features are colour coded according to their class: Shape (red), First Order (blue), GLCM (purple), GLRLM (orange), GLSZM (brown), GLDM (pink), and NGTDM (grey). Black stars and bold fonts indicate features where ICCs from original and Box–Cox transformed data are significantly different.

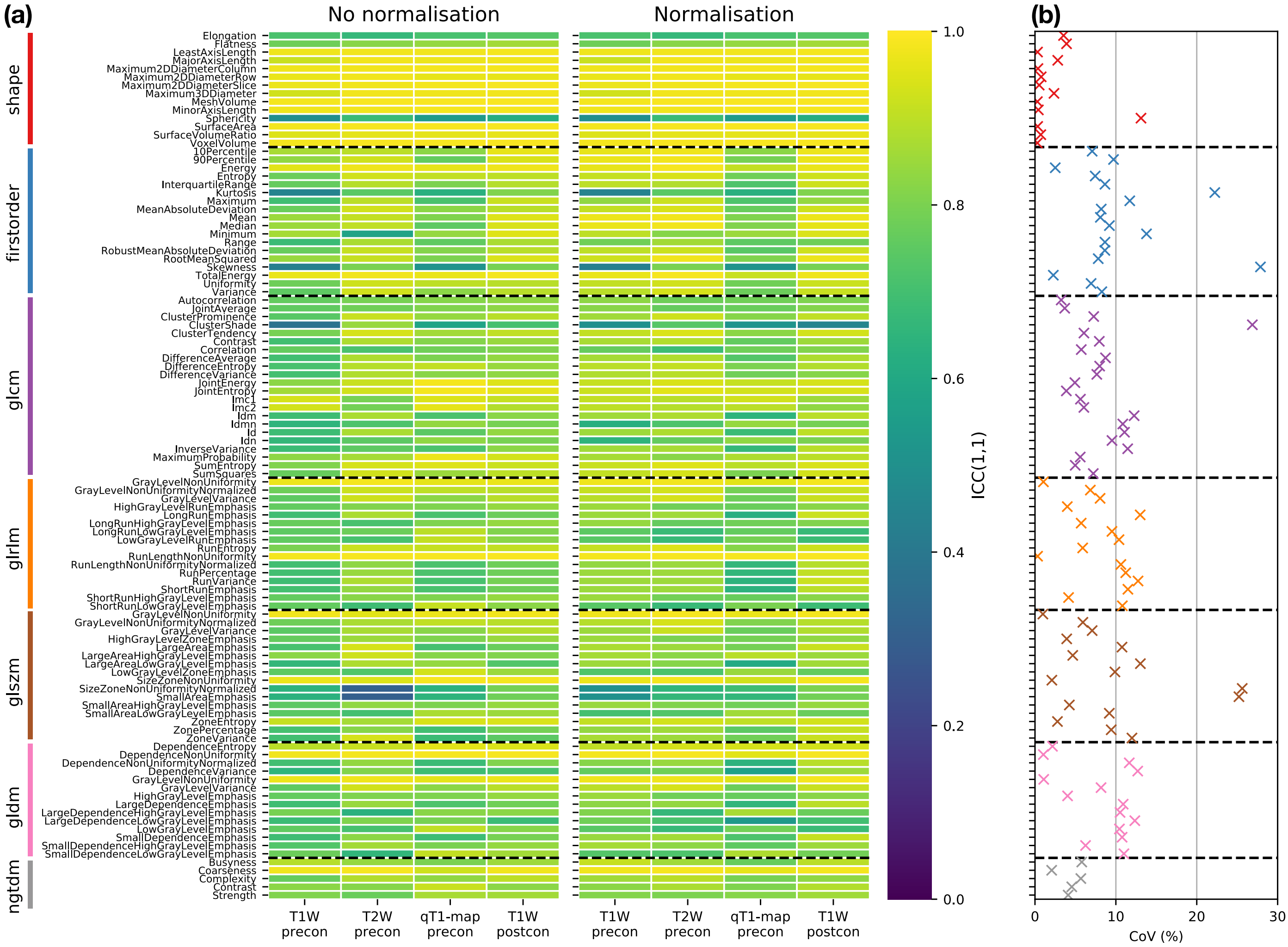

Figure 4. (**a**) ICC point estimates for all Box–Cox transformed features (rows), for each sequence (columns), from images without (left panel) and with (right panel) normalisation. Horizontal dashed black lines separate features in different classes. (**b**) Variability of ICCs across sequences and normalisations, plotted as the coefficient of variation (CoV) of ICCs over the eight datasets shown in the columns of (**a**). Features are colour-coded according to their class, and horizontal dashed black lines separate features in different classes.

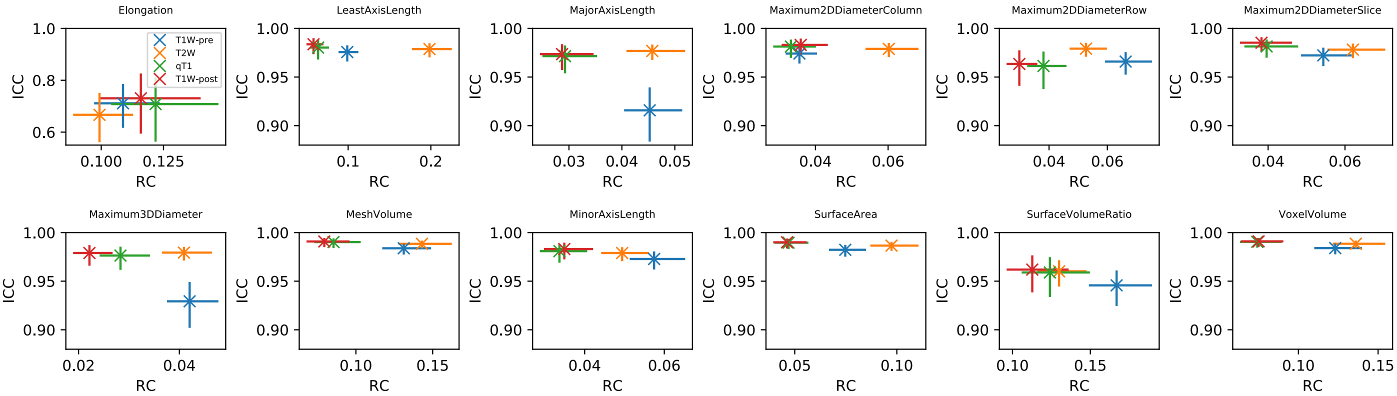

Figure 5. Plots of RC against ICC for Shape features, for four sequences (colours). In each panel, data points and error bars represent point estimates and 95% CIs.

As ICCs are more readily comparable across features and datasets, the following sections analyse ICCs in more detail, comparing non-normalised data across the different MR sequences, and investigating the effect of normalisation.

3.3. Comparing MR Sequences

Figures 6–8 plot ICCs for non-normalised images from the 4 MR sequences. Note that Figure 6 includes data from 134 lesions, while in Figures 7 and 8 the presented data come from the subset of 65 T_1W pre-contrast lesions that match those used in qT_1 map and T_1W post-contrast analyses. The bottom panel in Figures 6–8 plot the differences in ICCs between two sequences, allowing comparison of pre-contrast anatomical images (T_1W and T_2W), comparison of pre-contrast anatomical and quantitative images (T_1W and qT_1 map), and comparison of pre- and post-contrast images (T_1W pre- and post-contrast). These specific comparisons will be described below. As in Figure 3 above, note that these plots show the difference in ICCs after applying Fisher's *Z*-transformation, and the error bars show the 95% CI on this difference.

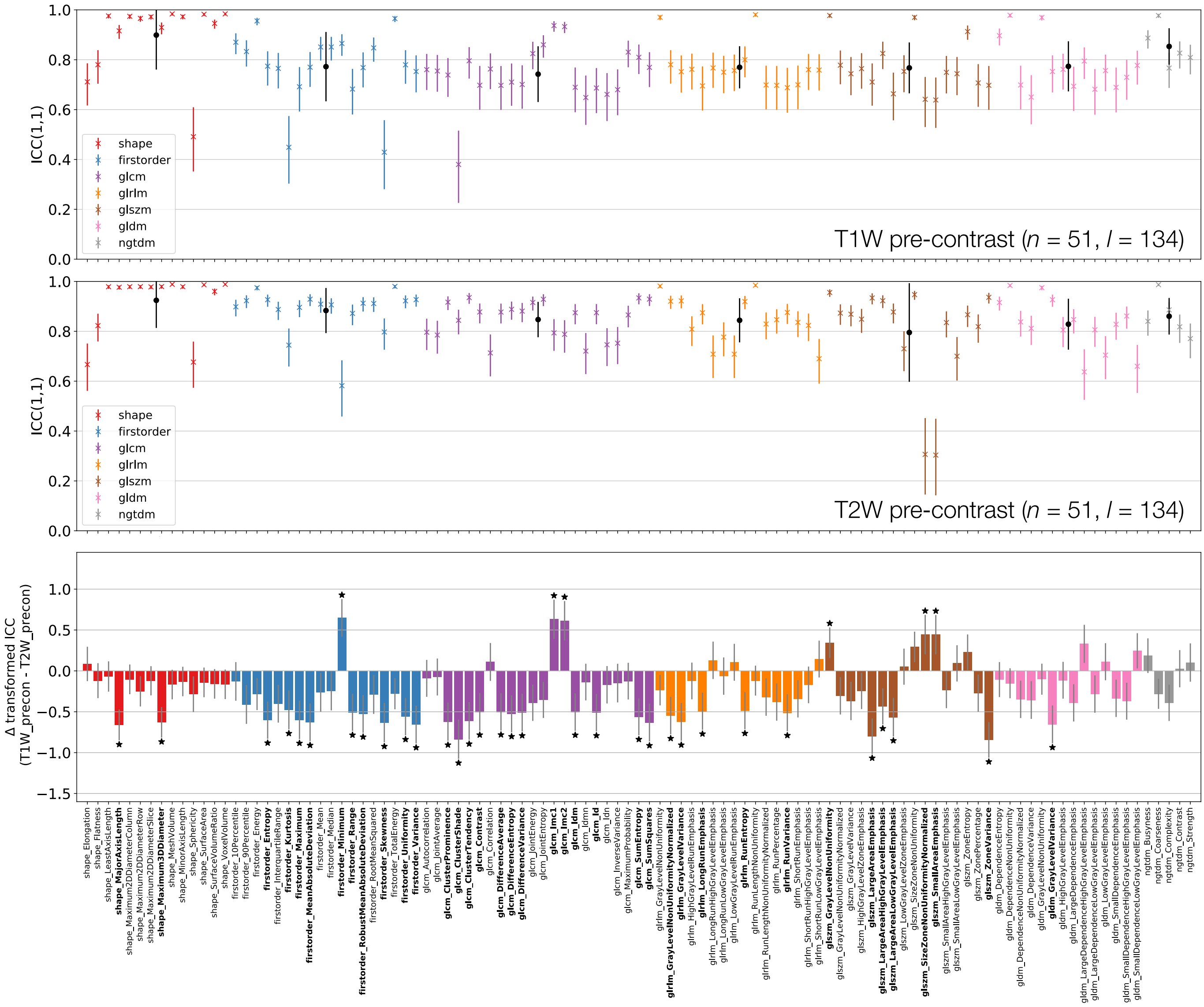

Figure 6. ICCs for Box–Cox transformed features from T_1W (top) and T_2W (middle) pre-contrast images. Data points and error bars represent ICC point estimates and 95% CIs. Features are colour coded according to their class and black points correspond to mean ± standard deviation ICCs over features within each class. The bottom panel represents the difference in ICCs (after applying Fisher's *Z*-transformation) for T_1W and T_2W pre-contrast images. Black stars and bold fonts indicate features where ICCs from the two sequences are significantly different.

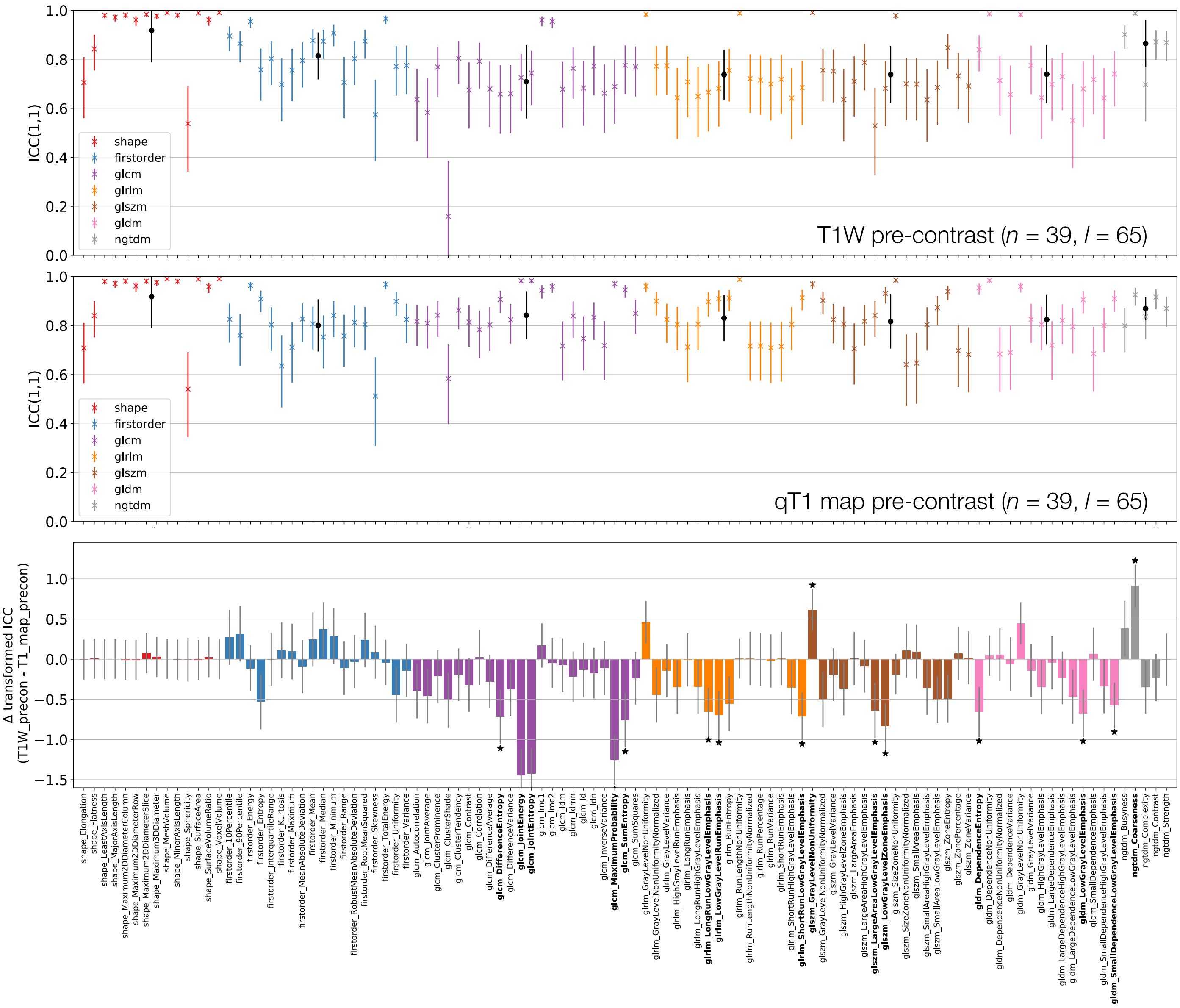

Figure 7. ICCs for Box–Cox transformed features from T_1W (top) and qT_1 map (middle) pre-contrast images. Data points and error bars represent ICC point estimates and 95% CIs. Features are colour coded according to their class and black points correspond to mean ± standard deviation ICCs over features within each class. The bottom panel represents the difference in ICCs (after applying Fisher's *Z*-transformation) for T_1W and qT_1 map pre-contrast images. Black stars and bold fonts indicate features where ICCs from the two sequences are significantly different.

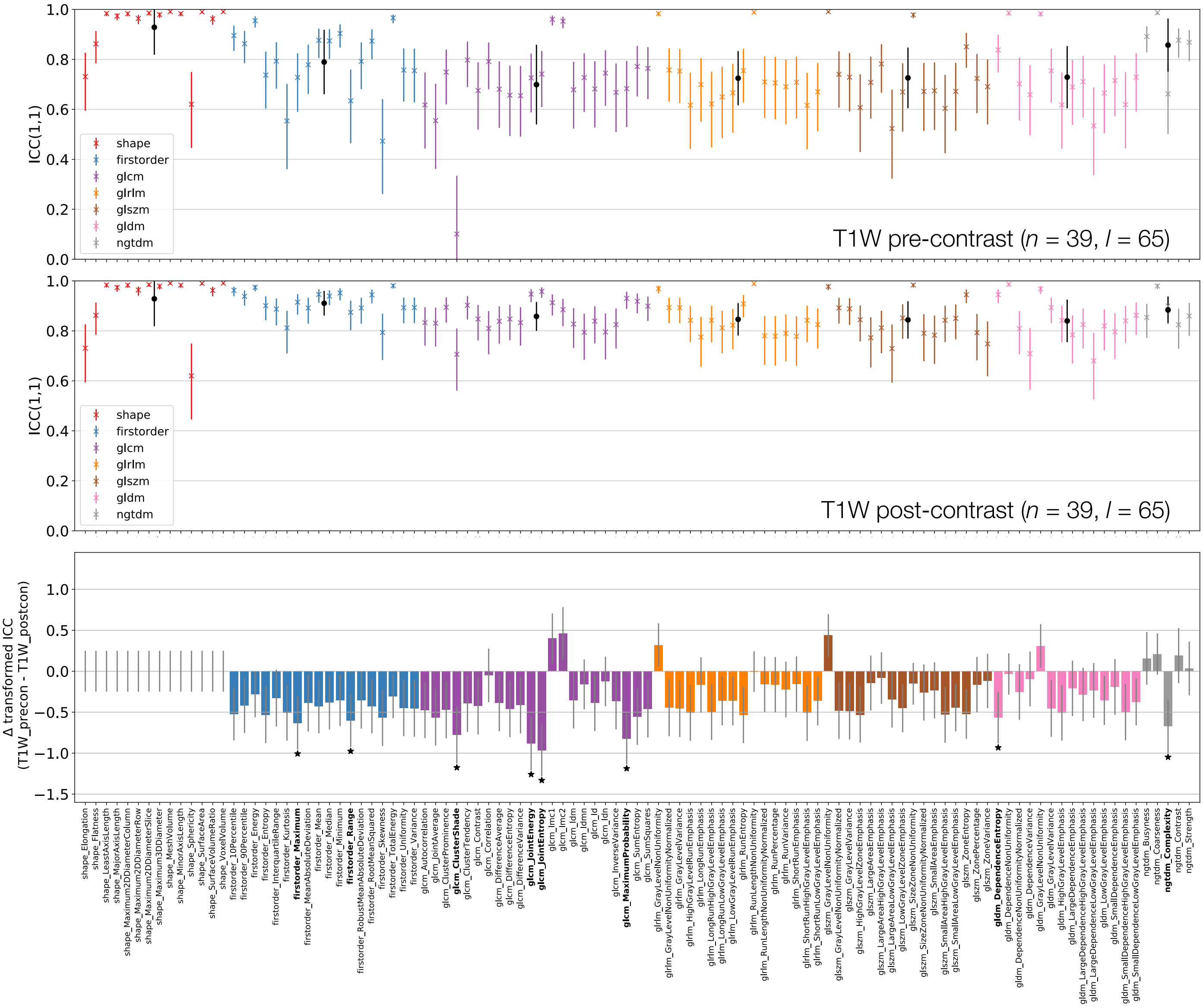

Figure 8. ICCs for Box–Cox transformed features from T_1W pre-contrast (top) and T_1W post-contrast (middle) images. Data points and error bars represent ICC point estimates and 95% CIs. Features are colour coded according to their class and black points correspond to mean ± ICCs over features within each class. The bottom panel represents the difference in ICCs (after applying Fisher's *Z*-transformation) for T_1W pre-contrast and T_1W post-contrast. Black stars and bold fonts indicate features where ICCs from the two sequences are significantly different.

3.3.1. T_1W Pre-Contrast and T_1W Pre-Contrast

Figure 6 shows that shape feature ICCs tend to be very similar between T_1W and T_2W pre-contrast images. Point estimates for Sphericity are most dissimilar, although the relatively wide confidence intervals means this difference is not significant; Major Axis Length and Maximum 3D Diameter ICCs are significantly higher on T_2W, though point estimates >0.91 on both images. Of the 38/105 features whose ICCs are significantly different, 10 are from the First Order class, with all but one of these (Minimum) exhibiting a higher ICC on T_2W. Of the remaining 26 significantly different features in the texture classes, 21 exhibit a higher ICC on T_2W.

3.3.2. T_1W Pre-Contrast and qT_1 map Pre-Contrast

Figure 7 shows that First Order features have comparable ICCs from T_1W pre-contrast images and qT_1 maps. Shape features, and therefore their associated ICCs, are essentially identical here as the same mask was used for both sets of images, with minor differences due to the different resolutions (see Section 2.1); note that qT_1 maps had a lower resolution than all other images. In all, 15/105 features have significantly different ICCs. Thirteen of these 15 show higher ICCs from qT_1 maps, and while most features are not significantly different, the tendency is for ICCs to be higher for qT_1 maps. Note that CIs tend to be wider on the T_1W pre-contrast plot here compared with the top row of Figure 6, as fewer lesions were included in the analysis for this comparison.

3.3.3. T_1W Pre- and Post-Contrast

Figure 8 shows that all First Order features from T_1W post-contrast images have higher ICCs than those from T_1W pre-contrast images, though only two are statistically significant. This tends to be true for all other classes, with the vast majority of features exhibiting higher ICCs on T_1W post-contrast images. Of the 8/105 features whose ICCs do differ significantly, all have ICCs which are higher on T_1W post-contrast images. Note that Shape features here are identical for pre- and post-contrast images, as the same masks were used for both.

3.4. Effect of Normalisation

Figures 9 and 10 show how ICCs are affected by applying image normalisation prior to feature extraction, comparing ICCs between images with and without normalisation, for each sequence. Normalisation tends to affect T_1W and T_2W pre-contrast images more than qT_1 maps or T_1W post-contrast images, with 30 and 14 ICCs significantly changed by applying normalisation to T_1W and T_2W pre-contrast images, respectively, with only seven ICCs significantly affected on qT_1 maps, and no ICCs significantly affected for T_1W post-contrast. For T_1W pre-contrast images, all ICCs which are significantly affected are higher when normalisation is applied; for T_2W, 11 of the 14 ICCs significantly affected improve with normalisation. Conversely, for qT_1 maps, although most ICCs are unaffected by normalisation, for the seven features which are significantly different, all have lower ICCs when normalisation is applied. Note that Shape features only depend on the masks, and so are unaffected by normalisation; their ICCs are therefore identical with and without normalisation.

3.5. Summary of Results

Taken together, the results from the present analysis highlight several aspects of radiomic feature repeatability which may be important to consider in future studies. Firstly, while most features had non-Gaussian distributions, the use of Box–Cox transformations enabled ICCs and RCs to be calculated appropriately for an average of 97% of features across sequences. Secondly, features exhibited a wide range of ICCs, with Shape features tending to have the highest ICCs. Thirdly, 19% of features from non-normalised images exhibited significantly different ICCs in pair-wise comparisons between different MR acquisitions. Fourthly, the use of image normalisation tended to increase ICCs for pre-contrast T_1- and T_2-weighted images, and decrease ICCs for qT_1 maps. Finally, RCs and ICCs can provide different insights into feature repeatability.

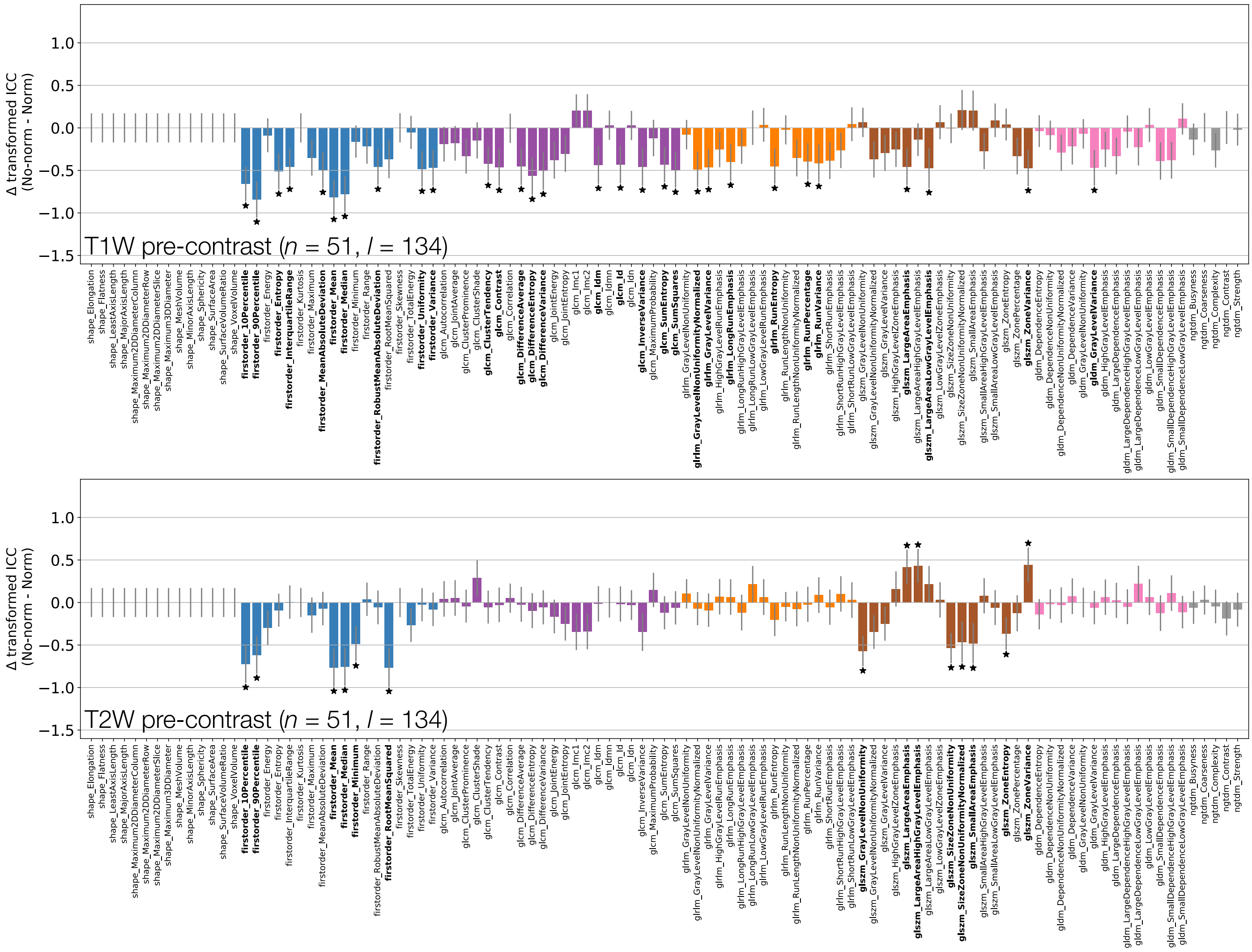

Figure 9. Effect of image normalisation on T_1W and T_2W pre-contrast images. Comparison of ICCs for Box–Cox transformed features from non-normalised and normalised images, for T_1W (top) and T_2W (bottom) pre-contrast images. Bars represent the difference in ICC point estimates (after applying Fisher's *Z*-transformation), and error bars represent 95% CIs. Features are colour coded according to their class. Black stars and bold fonts indicate features where ICCs from non-normalised and normalised images are significantly different.

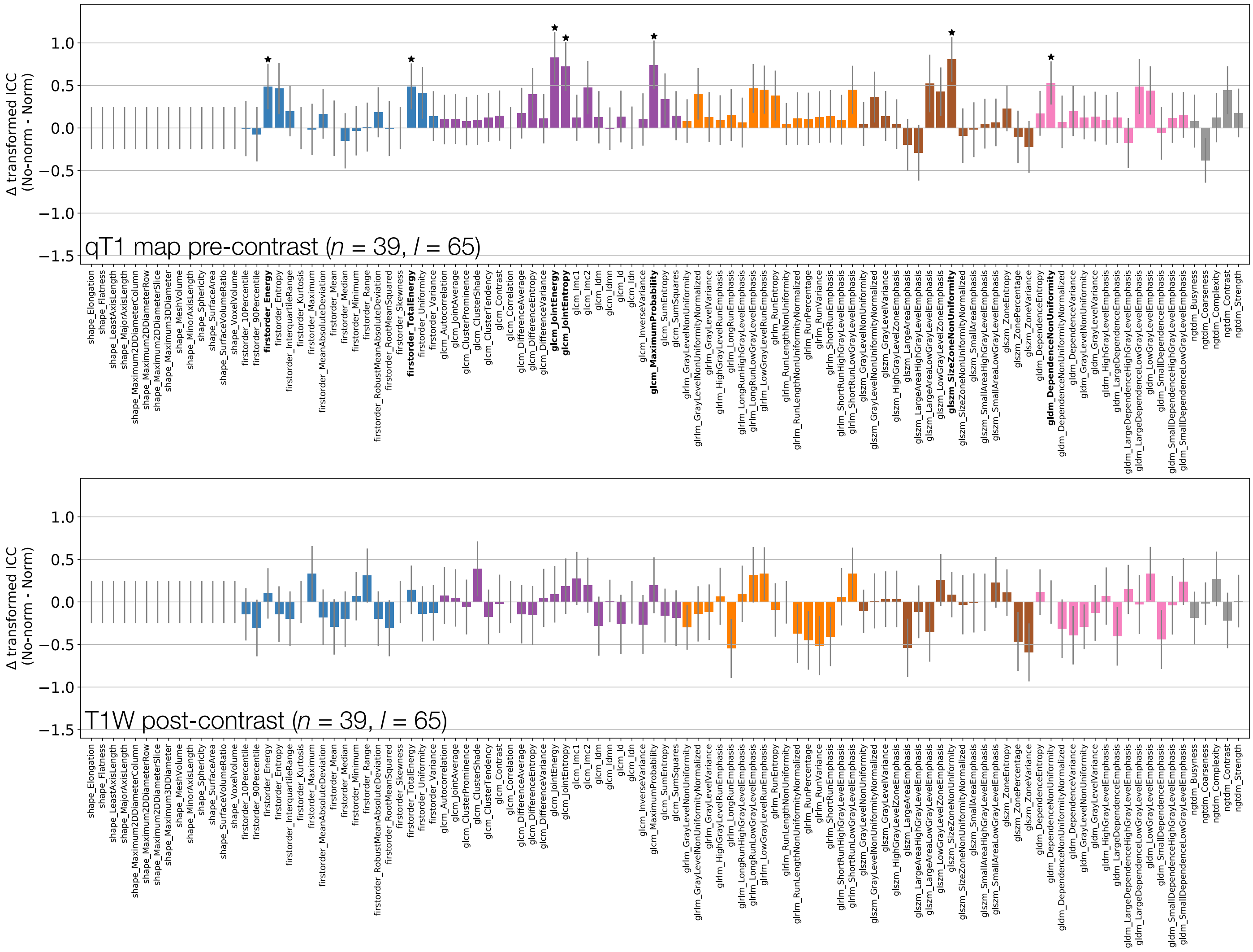

Figure 10. Effect of image normalisation on qT_1 maps and T_1W post-contrast images. Comparison of ICCs for Box–Cox transformed features from non-normalised and normalised images, for qT_1 map pre-contrast (top) and T_1W post-contrast (bottom). Bars represent the difference in ICC point estimates (after applying Fisher's *Z*-transformation), and error bars represent 95% CIs. Features are colour coded according to their class. Black stars and bold fonts indicate features where ICCs from non-normalised and normalised images are significantly different.

4. Discussion

Evaluating repeatability is a key step in the technical validation of imaging biomarkers, which itself is essential for translating such biomarkers into clinical practice [4]. In the context of radiomics, repeatability is an important factor when determining which features should be included in a predictive model. For example, if a feature has poor single site repeatability it is unlikely that it will have good multi-centre reproducibility, limiting its utility in a model. Conversely, good single site repeatability can be seen as a necessary, but not sufficient, condition for utility, as multi-centre reproducibility would still need to be demonstrated. As such, feature repeatability is a prerequisite for contributing to a robust predictive signature, or use as a biomarker of treatment response. Importantly, a feature's repeatability according to a particular metric may suggest its suitability for a particular use case, with a high ICC implying good performance when used as a diagnostic and/or predictive biomarker, whereas a low RC would be required when using a feature as a biomarker of treatment response. Note that what is considered 'low' in this context is dependent on the magnitude of the expected biomarker change, which needs to be considered alongside RC when evaluating the relative utility of features as biomarkers of change. Appropriately evaluating feature repeatability has practical consequences, as sample sizes and statistical power can be influenced by the repeatability of the biomarker; for example, a study using as a biomarker a feature with low repeatability would require a larger sample size than if a feature with high repeatability was used [28]. By assessing MR radiomic feature repeatability using two different metrics in a relatively large clinical cohort, investigating the effects of MR sequence, image normalisation, and assumptions about feature distributions, this work contributes to the technical validation of radiomic features. By focussing on liver metastases, and using quantitative T_1 maps and post-contrast T_1W images, this work complements existing repeatability studies using other MR sequences in other tumour types [12,14–20].

Evaluating repeatability using the metrics described in this work requires features to follow a Gaussian distribution, though this assumption is often not confirmed. This work found most radiomic feature distributions to be non-Gaussian, questioning the appropriateness of directly applying ICC or RC calculations. While ICCs are relatively robust to this assumption being invalid, it can make a significant difference for some radiomic features. As such, we suggest that examining feature distributions should form part of radiomic analyses, with results here demonstrating that Box–Cox transformations can be an effective way of obtaining Gaussian feature distributions. It is important to note that performing the Box–Cox transformation will alter some feature repeatability metrics significantly.

Across all datasets, Voxel Volume and Mesh Volume were the most repeatable features. As a class, Shape features tended to have the highest ICCs, consistent with observations on T_2W images of cervical tumours [15], quantitative diffusion kurtosis maps of prostate tumours [20], and quantitative apparent diffusion coefficient maps of liver metastases and ovarian tumours [19]. For T_2W images in rectal cancer, Gourtsoyianni et al. note that Gray Level Size Zone Matrix and Neighbouring Gray Tone Difference Matrix features tended to have poor repeatability [12]. In the present work, the features with the lowest ICCs on T_2W were from the Gray Level Size Zone Matrix, while Neighbouring Gray Tone Difference Matrix features performed reasonably well (ICCs > 0.77). Gourtsoyianni et al. did specifically note that Coarseness was an exception in terms of Neighbouring Gray Tone Difference Matrix features, which is consistent with the present work where it had ICC > 0.93 for all sequences. Of the nine non-shape features with ICC > 0.90 across all datasets in the present work, four (Energy, Total Energy, Run Length Non Uniformity, and Coarseness) were also found to have excellent repeatability and reproducibility in a phantom study using T_2W images [36], providing further evidence of their robustness; note that in the phantom study these features tended to come from images filtered prior to feature extraction, while filtering was not investigated in the present work. Along with differences in filtering, the variation between studies in terms of MR sequence, image normalisation, and tumour type, make it challenging to directly compare repeatabilities across studies. Even comparing across the same type of MR sequence can be confounded by different studies using different parameters, with radiomic features showing sensitivity to echo time and repetition time in T_2W acquisitions [36]. In the present work, the use of zero-filling during image reconstruction should also be noted, as this will tend to reduce intra-lesion heterogeneity in signal intensities, and hence impact many radiomic features. As the use of zero-filling differed between acquisitions, this will contribute to feature differences across sequences, in addition to the inherent MR weighting. Although not considered here, the reconstructed images could be resampled to achieve isotropic voxel sizes, which would be expected to impact texture features extracted in 3D. These points should also be noted in relation to multi-centre and multi-vendor reproducibility assessments, as precise acquisition and reconstruction details may vary between scanners, potentially impacting radiomic features. Further work is needed to understand the benefits of harmonising MR acquisitions and reconstructions for improving radiomic feature reproducibility, relative to post-acquisition harmonisation approaches [37]. Such approaches are especially relevant for retrospective studies where prospective acquisition harmonisation is not possible, and include the use of neural networks for pre-processing acquired images prior to feature extraction [37], and methods such as ComBat for mitigating the effects of feature variability related to specific centres or scanners [38].

Also note that in the present study, the nine non-shape features with ICC > 0.90 all showed strong correlations with Mesh Volume, which may contribute to their repeatability. Correlations between radiomic features and tumour volume have been reported previously [39]; indeed, it has been noted that using repeatability to guide feature selection may result in radiomic signatures which essentially reflect tumour volume [39]. As such, care must be taken to evaluate repeatability along with feature correlations.

In addition to their generally high ICCs, Shape features also tended to show the lowest variability in ICCs across sequences; this is to be expected as Shape features are insensitive to normalisation, and the same masks were used for T_1-weighted images and qT_1 maps. For a more comprehensive evaluation of Shape features, inter-observer variability in contouring could also be evaluated, and compared with test-retest repeatability; this would provide analogous data to that presented perviously for T_2W images of cervical tumours [15], where Shape features had high ICCs for both test-retest repeatability (all but one feature having ICC > 0.9) and inter-observer reproducibility (all features having ICC > 0.9). The low inter-sequence variability in Shape ICCs observed here may also imply that Shape feature repeatability does not strongly depend on MR image contrast; note however that RC values for Shape features tended to show more variation across sequences, implying that the contrast can impact repeatability. (This effect of contrast can only stem from the fact that tumours may be more readily distinguished from surrounding tissue on some sequences, facilitating more repeatable delineation; once contours are defined, Shape features are independent of signal intensities). In general, ICCs and RCs provide different information about feature repeatability, and the choice of metric needs to be considered in a study-specific context. Practically, this means that a given feature from a given sequence may have favourable repeatability characteristics for one application (for example, as a diagnostic or predictive biomarker), but less favourable characteristics for another (for example, as a biomarker of treatment response).

The ICCs of non-shape features tended to exhibit greater sensitivity to sequence and normalisation, particularly for Skewness, Cluster Shade, Small Area Emphasis and Size Zone Non-Uniformity Normalized. As such, it may be expected that the repeatability of such features may vary more across different studies, if different MR acquisitions and normalisation approaches are used, necessitating study-specific repeatability analyses. Furthermore, unlike for RC, a feature's ICC depends on the inter-subject variability of that feature [28], meaning that repeatability as evaluated from ICCs may vary across different cohorts, even for the same sequence and normalisation, further motivating repeatability assessments on a per-study basis using the most appropriate metric for that study.

When comparing ICCs from non-normalised T_1W and T_2W pre-contrast images, First Order features tended to show better repeatability on T_2W. This could reflect easier lesion definition on T_2W images, leading to more repeatable signal intensity distributions, or could be due to T_1W signals having sensitivity to genuine changes in the lesions between scans, resulting in apparently poorer repeatability. Normalisation tended to improve repeatability for both sequences, with ICCs increasing significantly in 29% and 13% of features for T_1W and T_2W, respectively. This improvement with normalisation would be expected if repeated scans differed by a uniform scaling factor. However, it should not be assumed that this is the only difference, as non-uniformities may be present that could vary between visits, for example through B_0 or B_1 inhomogeneities [40]. Although the differences are not dramatic, these ICC results suggest a slight preference of T_2W over T_1W images for obtaining repeatable radiomic features in this dataset.

As image signal intensities have no inherent units in MR, it cannot be assumed that they can be directly compared across different sequences, or even across repeated acquisitions of the same sequence. A key motivation in the use of quantitative MR techniques is their ability to yield biomarkers which are independent of absolute image signal intensities. However, comparing First Order features between non-normalised T_1W and qT_1 pre-contrast images showed no significant differences in repeatability, implying that signal intensities are as comparable as quantitative T_1 values; note that this must be considered in the context of the present study, where a single scanner was used, with the same acquisitions performed for all repeat scans. While quantitative values may be expected to be more repeatable than signal intensities, quantitative maps may be more sensitive to motion, given that T_1 is here quantified through modelling signals across three separate acquisitions. Radiomic features will in general be affected by motion artefacts, which are especially relevant for free-breathing liver acquisitions, and these effects may differ between weighted and quantitative images. As applying normalisation had relatively little effect on repeatabilities from qT_1 maps, but tended to improve those from T_1W images, the use of normalisation resulted in T_1W images yielding higher ICCs than qT_1 maps for all First Order features, with eight of these being significantly higher. This highlights the different effects of normalisation on weighted and quantitative images. It could be argued that normalisation is more appropriate for weighted images, as it aims to correct for potential scaling differences between signal intensities, while quantitative maps are insensitive to such scalings. Contrary to this, it has been reported that normalisation improved repeatability for quantitative maps (the apparent diffusion coefficient), but not for weighted images (T_2W) [17]. Further work is therefore needed to fully understand the use of quantitative maps, as opposed to weighted images, as radiomics inputs. This may be especially important for multi-centre studies, where comparison of signal intensities across scanners is likely to be more problematic than for a single scanner, and quantitative maps may be expected to yield more reproducible features. Although only a simple linear normalisation has been used in the present work, it should be noted that there are many approaches that could be employed [41–44]. The issue of normalisation is also especially relevant when seeking to compare radiomic features over time, for example in assessing treatment response; here, the use of normalisation could confound genuine signal intensity changes. Given the conflicting results and importance of this step, further investigation of normalisation techniques, along with the use of quantitative maps, is warranted. The potential for using normalised voxel sizes could also be investigated, and may be especially relevant for comparisons between anatomical images and quantitative maps, given that the former would typically be acquired at a higher resolution.

Although most differences are not statistically significant, there is a tendency for higher ICCs on T_1W post-contrast images than on T_1W pre-contrast images. This cannot be explained by improved lesion conspicuity post-contrast, as the same masks were used for both sequences and were drawn on pre-contrast images. The trend for better repeatability could be related to a higher signal-to-noise ratio post-contrast, due to T_1 shortening, or could reflect an increased inter-subject variability due to differences in contrast agent uptake. Of the four sequences assessed here, T_1W post-contrast features were least affected by normalisation, with no features showing significant differences. Note that throughout this study, statistical significance was judged based on a Bonferroni-adjusted p-value which accounted for the dominant source of multiple comparisons, namely, the number of features extracted. Given the exploratory nature of the investigation, further adjustments for comparisons between different sequences were not included; these could be incorporated into future prospective comparison studies to lower the threshold for determining significance.

The main limitations of the study are that repeatability was assessed over a period of 2–7 days, and that motion correction techniques were not applied. By having up to a week between repeat scans for some patients, it is possible that tumours may have undergone genuine macroscopic or microscopic changes, which radiomic features may be sensitive to. This could lead to an underestimate of feature repeatability, and could be mitigated by performing test-retest studies with features evaluated from scans performed minutes apart [15,16,20]; note however that this would not be feasible for T_1W post-contrast scans, as the contrast agent administration cannot be repeated in such a short time window. As liver motion could affect the comparison between different sequences, future work could investigate the impact of applying motion correction before extracting radiomic features from multiple sequences. Finally, in the context of biomarker validation, multi-centre and multi-vendor reproducibility is an essential next step. Comparison across scanners may be especially important for radiomic features which directly depend on signal intensities, as these are unlikely to be directly comparable across scanners. While the use of quantitative parametric maps removes the dependence on signal intensities, quantitative T_1 values in the human brain have been reported to differ between vendors [45], indicating that there is also a need for inter-vendor comparisons of radiomic features from quantitative maps.

5. Conclusions

Radiomic features from colorectal cancer liver metastases exhibit a wide range of repeatabilities. Some, but not all, of these can be significantly affected by the choice of MR sequence, image quantitation, use of contrast agent, and image normalisation. The presence, magnitude and direction of significant changes due to these factors is not readily predictable without conducting such an analysis in a test cohort, as performed here for one study scenario. The choice of repeatability metric can influence conclusions regarding feature repeatability, and the study context should be used to determine the most appropriate metric. These general principles are likely to extend to other MR studies using different sequences, suggesting that feature-specific repeatability, from specific combinations of MR sequence and pre-processing steps, should be evaluated with the most appropriate metric in order to select robust radiomic features as biomarkers in specific studies.

Supplementary Materials: The following are available online at https://www.mdpi.com/2072-6694/13/2/240/s1, Figure S1, example masks for T1W pre-contrast, qT1 map pre-contrast, and T1W pre-contrast subset; Figures S2–5, feature histograms, Box-Cox normality plots, transformed feature histograms, and QQ-plots for T1W pre-contrast, T2 pre-contrast, qT1 map pre-contrast, and T1W post-contrast; Figure S6, effect of applying Box-Cox transforms to feature distributions, for qT1 map pre-contrast and T1W post-contrast; Figures S7–14, Bland-Altman plots with RC values for all features and datasets; Figures S15–20, plots of RC against ICC for each feature class except Shape.

Author Contributions: Conceptualization, J.P.B.O.; methodology, D.J.M., N.P., G.J.M.P. and J.P.B.O.; software, D.J.M. and S.C.; validation, D.J.M., R.A.L., G.J.M.P. and J.P.B.O.; formal analysis, D.J.M., N.P., S.C., and J.P.B.O.; investigation, D.J.M. and J.P.B.O.; data curation, D.J.M., R.A.L., Y.W., G.J.M.P., G.C.J., and J.P.B.O.; writing—original draft preparation, D.J.M. and J.P.B.O.; writing—review and editing, all authors; visualization, D.J.M.; supervision, J.P.B.O.; funding acquisition, G.J.M.P., G.C.J. and J.P.B.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Cancer Research UK Accelerator Award C19221/A28683 (to JOC), Cancer Research UK Advanced Clinician Scientist Fellowship C19221/A22746 (to JOC), and Cancer Research UK Manchester Cancer Research Centre C147/A18083 (to GJ). This study represents independent research supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and the Institute of Cancer Research, London and by the Manchester NIHR Biomedical Research Centre. Original MRI data were acquired through an unrestricted grant from Roche. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the North West - Greater Manchester Central Research Ethics Committee (REC 09/H1008/99), start date 25 June 2009.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Extracted radiomic features can be found at https://gitlab.com/manchester_qbi/manchester_qbi_public/radiomics_repeatability.

Acknowledgments: The authors thank Hitesh B. Mistry and Martin A. Stoffel for useful discussions.

Conflicts of Interest: Geoff J.M. Parker has a shareholding and part time appointment and directorship at Bioxydyn Ltd. All other authors declare no competing interests.

References

1. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; Van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446. [CrossRef] [PubMed]
2. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Cavalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006. [CrossRef] [PubMed]
3. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577. [CrossRef] [PubMed]
4. O'Connor, J.P.B.; Aboagye, E.O.; Adams, J.E.; Aerts, H.J.W.L.; Barrington, S.F.; Beer, A.J.; Boellaard, R.; Bohndiek, S.E.; Brady, M.; Brown, G.; et al. Imaging biomarker roadmap for cancer studies. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 169–186. [CrossRef]
5. O'Connor, J.P.B. Rethinking the role of clinical imaging. *eLife* **2017**, *6*, e30563. [CrossRef]
6. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; De Jong, E.E.C.; Van Timmeren, J.; Sanduleanu, S.; Larue, R.T.H.M.; Even, A.J.G.; Jochems, A.; et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762. [CrossRef]
7. Fornacon-Wood, I.; Faivre-Finn, C.; O'Connor, J.P.B.; Price, G.J. Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. *Lung Cancer* **2020**, *146*, 197–208. [CrossRef]
8. Sun, R.; Limkin, E.J.; Vakalopoulou, M.; Dercle, L.; Champiat, S.; Han, S.R.; Verlingue, L.; Brandao, D.; Lancia, A.; Ammari, S.; et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: An imaging biomarker, retrospective multicohort study. *Lancet Oncol.* **2018**, *19*, 1180–1191. [CrossRef]
9. Sullivan, D.C.; Obuchowski, N.A.; Kessler, L.G.; Raunig, D.L.; Gatsonis, C.; Huang, E.P.; Kondratovich, M.; McShane, L.M.; Reeves, A.P.; Barboriak, D.P.; et al. Metrology Standards for Quantitative Imaging Biomarkers. *Radiology* **2015**, *277*, 813–825. [CrossRef]
10. Fournier, L.; Costaridou, L.; Bidaut, L.; Michoux, N.; Lecouvet, F.; de Geus-Oei, L.; Boellaard, R.; Oprea-Lager, D.E.; Obuchowski, N.; Caroli, A.; et al. Incorporating radiomics into clinical trials: Expert consensus on considerations for data-driven compared to biologically-driven quantitative biomarkers. *Eur. Radiol.* **2020**, in press.
11. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* **2018**, *102*, 1143–1158. [CrossRef] [PubMed]
12. Gourtsoyianni, S.; Doumou, G.; Prezzi, D.; Taylor, B.; Stirling, J.J.; Taylor, N.J.; Siddique, M.; Cook, G.J.R.; Glynne-Jones, R.; Goh, V. Primary rectal cancer: Repeatability of global and local-regional MR imaging texture features. *Radiology* **2017**, *284*, 552–561. [CrossRef] [PubMed]
13. Kickingereder, P.; Neuberger, U.; Bonekamp, D.; Piechotta, P.L.; Götz, M.; Wick, A.; Sill, M.; Kratz, A.; Shinohara, R.T.; Jones, D.T.W.; et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro-Oncology* **2018**, *20*, 848–857. [CrossRef] [PubMed]
14. Reischauer, C.; Patzwahl, R.; Koh, D.M.; Froehlich, J.M.; Gutzeit, A. Texture analysis of apparent diffusion coefficient maps for treatment response assessment in prostate cancer bone metastases—A pilot study. *Eur. J. Radiol.* **2018**, *101*, 184–190. [CrossRef]
15. Fiset, S.; Welch, M.L.; Weiss, J.; Pintilie, M.; Conway, J.L.; Milosevic, M.; Fyles, A.; Traverso, A.; Jaffray, D.; Metser, U.; et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiother. Oncol.* **2019**, *135*, 107–114. [CrossRef]
16. Mahon, R.N.; Hugo, G.D.; Weiss, E. Repeatability of texture features derived from magnetic resonance and computed tomography imaging and use in predictive models for non-small cell lung cancer outcome. *Phys. Med. Biol.* **2019**, *64*, 145007. [CrossRef]
17. Schwier, M.; van Griethuysen, J.; Vangel, M.G.; Pieper, S.; Peled, S.; Tempany, C.; Aerts, H.J.W.L.; Kikinis, R.; Fennessy, F.M.; Fedorov, A. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci. Rep.* **2019**, *9*, 9441. [CrossRef]
18. Barrett, T.; Lawrence, E.M.; Priest, A.N.; Warren, A.Y.; Gnanapragasam, V.J.; Gallagher, F.A.; Sala, E. Repeatability of diffusion-weighted MRI of the prostate using whole lesion ADC values, skew and histogram analysis. *Eur. J. Radiol.* **2019**, *110*, 22–29. [CrossRef]
19. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; Ibrahim, A.; Van Beers, B.E.; Heerschap, A.; Jackson, A.; Wildberger, J.E.; Mottaghy, F.M.; DeSouza, N.M.; et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci. Rep.* **2019**, *9*, 4800. [CrossRef]
20. Merisaari, H.; Taimen, P.; Shiradkar, R.; Ettala, O.; Pesola, M.; Saunavaara, J.; Boström, P.J.; Madabhushi, A.; Aronen, H.J.; Jambor, I. Repeatability of radiomics and machine learning for DWI: Short-term repeatability study of 112 patients with prostate cancer. *Magn. Reson. Med.* **2020**, *83*, 2293–2309. [CrossRef]
21. Shiri, I.; Hajianfar, G.; Sohrabi, A.; Abdollahi, H.P.; Shayesteh, S.; Geramifar, P.; Zaidi, H.; Oveisi, M.; Rahmim, A. Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: Test–retest and image registration analyses. *Med. Phys.* **2020**, *47*, 4265–4280. [CrossRef] [PubMed]

22. Fiz, F.; Viganò, L.; Gennaro, N.; Costa, G.; La Bella, L.; Boichuk, A.; Cavinato, L.; Sollini, M.; Politi, L.S.; Chiti, A.; et al. Radiomics of liver metastases: A systematic review. *Cancers* **2020**, *12*, 2881. [CrossRef]
23. Jayson, G.C.; Zhou, C.; Backen, A.; Horsley, L.; Marti-Marti, K.; Shaw, D.; Mescallado, N.; Clamp, A.; Saunders, M.P.; Valle, J.W.; et al. Plasma Tie2 is a tumor vascular response biomarker for VEGF inhibitors in metastatic colorectal cancer. *Nat. Commun.* **2018**, *9*, 4672. [CrossRef] [PubMed]
24. Fram, E.K.; Herfkens, R.J.; Johnson, G.A.; Glover, G.H.; Karis, J.P.; Shimakawa, A.; Perkins, T.G.; Pelc, N.J. Rapid calculation of T1 using variable flip angle gradient refocused imaging. *Magn. Reson. Imaging* **1987**, *5*, 201–208. [CrossRef]
25. Van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J.W.L. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
26. Fornacon-Wood, I.; Mistry, H.; Ackermann, C.J.; Blackhall, F.; McPartlin, A.; Faivre-Finn, C.; Price, G.J.; O'Connor, J.P.B. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *Eur. Radiol.* **2020**, *30*, 6241–6250. [CrossRef]
27. McGraw, K.O.; Wong, S.P. Forming Inferences about Some Intraclass Correlation Coefficients. *Psychol. Methods* **1996**, *1*, 30–46. [CrossRef]
28. Weir, J.P. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Cond. Res.* **2005**, *19*, 231–240.
29. Liljequist, D.; Elfving, B.; Roaldsen, K.S. Intraclass correlation—A discussion and demonstration of basic features. *PLoS ONE* **2019**, *14*, e0219854. [CrossRef]
30. Barnhart, H.X.; Barboriak, D.P. Applications of the repeatability of quantitative imaging biomarkers: A review of statistical analysis of repeat data sets. *Transl. Oncol.* **2009**, *2*, 231–235. [CrossRef]
31. Raunig, D.L.; McShane, L.M.; Pennello, G.; Gatsonis, C.; Carson, P.L.; Voyvodic, J.T.; Wahl, R.L.; Kurland, B.F.; Schwarz, A.J.; Gönen, M.; et al. Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment. *Stat. Methods Med. Res.* **2015**, *24*, 27–67. [CrossRef] [PubMed]
32. Ionan, A.C.; Polley, M.Y.C.; McShane, L.M.; Dobbin, K.K. Comparison of confidence interval methods for an intra-class correlation coefficient (ICC). *BMC Med Res. Methodol.* **2014**, *14*, 121. [CrossRef] [PubMed]
33. Box, G.E.P.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. B* **1964**, *26*, 211–252. [CrossRef]
34. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
35. Donner, A.; Zou, G. Testing the equality of dependent intraclass correlation coefficient. *Statistician* **2002**, *51*, 367–379. [CrossRef]
36. Bianchini, L.; Santinha, J.; Loução, N.; Figueiredo, M.; Botta, F.; Origgi, D.; Cremonesi, M.; Cassano, E.; Papanikolaou, N.; Lascialfari, A. 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. *Magn. Reson. Med.* **2020**, *85*, 1713–1726. [CrossRef]
37. Da-ano, R.; Visvikis, D.; Hatt, M. Harmonization strategies for multicenter radiomics investigations. *Phys. Med. Biol.* **2020**. [CrossRef]
38. Da-ano, R.; Masson, I.; Lucia, F.; Doré, M.; Robin, P.; Alfieri, J.; Rousseau, C.; Mervoyer, A.; Reinhold, C.; Castelli, J.; et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci. Rep.* **2020**, *10*, 10248. [CrossRef]
39. Welch, M.L.; McIntosh, C.; Haibe-Kains, B.; Milosevic, M.F.; Wee, L.; Dekker, A.; Huang, S.H.; Purdie, T.G.; O'Sullivan, B.; Aerts, H.J.W.L.; et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother. Oncol.* **2019**, *130*, 2–9. [CrossRef]
40. Belaroussi, B.; Milles, J.; Carme, S.; Zhu, Y.M.; Benoit-Cattin, H. Intensity non-uniformity correction in MRI: Existing methods and their validation. *Med Image Anal.* **2006**, *10*, 234–246. [CrossRef]
41. Nyúl, L.G.; Udupa, J.K. On standardizing the MR image intensity scale. *Magn. Reson. Med.* **1999**, *42*, 1072–1081. [CrossRef]
42. Nyúl, L.G.; Udupa, J.K.; Zhang, X. New variants of a method of MRI scale standardization. *IEEE Trans. Med Imaging* **2000**, *19*, 143–150. [CrossRef] [PubMed]
43. Shah, M.; Xiao, Y.; Subbanna, N.; Francis, S.; Arnold, D.L.; Collins, D.L.; Arbel, T. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Med. Image Anal.* **2011**, *15*, 267–282. [CrossRef] [PubMed]
44. Shinohara, R.T.; Sweeney, E.M.; Goldsmith, J.; Shiee, N.; Mateen, F.J.; Calabresi, P.A.; Jarso, S.; Pham, D.L.; Reich, D.S.; Crainiceanu, C.M. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin.* **2014**, *6*, 9–19. [CrossRef]
45. Lee, Y.; Callaghan, M.F.; Acosta-Cabronero, J.; Lutti, A.; Nagy, Z. Establishing intra- and inter-vendor reproducibility of T1 relaxation time measurements with 3T MRI. *Magn. Reson. Med.* **2019**, *81*, 454–465. [CrossRef]