Could normalization improve robustness of abdominal MRI radiomic features?

PAPER • **OPEN ACCESS**

Could normalization improve robustness of abdominal MRI radiomic features?

To cite this article: Valentina Giannini *et al* 2023 *Biomed. Phys. Eng. Express* **9** 055002

Biomedical Physics & Engineering Express

PAPER

Could normalization improve robustness of abdominal MRI radiomic features?

Valentina Giannini[1,2], Jovana Panic[2,3], Daniele Regge[2], Gabriella Balestra[3] and Samanta Rosati[3]

[1] University of Turin, Department of Surgical Science, Turin, Italy
[2] Candiolo Cancer Institute, FPO-IRCCS, Candiolo (TO), Italy
[3] Polytechnic of Turin, Department of Electronics and Telecommunications, Turin, Italy

E-mail: valentina.giannini@unito.it

Keywords: image normalization, feature normalization, abdominal MRI radiomics, multi-center dataset, feature extraction

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
22 December 2022

REVISED
23 June 2023

ACCEPTED FOR PUBLICATION
6 July 2023

PUBLISHED
17 July 2023

Abstract

Radiomics-based systems could improve the management of oncological patients by supporting cancer diagnosis, treatment planning, and response assessment. However, one of the main limitations of these systems is the generalizability and reproducibility of results when they are applied to images acquired in different hospitals by different scanners. Normalization has been introduced to mitigate this issue, and two main approaches have been proposed: one rescales the image intensities (*image normalization*), the other the feature distributions for each center (*feature normalization*). The aim of this study is to evaluate how different image and feature normalization methods impact the robustness of 93 radiomics features acquired using a multicenter and multi-scanner abdominal Magnetic Resonance Imaging (MRI) dataset. To this scope, 88 rectal MRIs were retrospectively collected from 3 different institutions (4 scanners), and for each patient, six 3D regions of interest on the obturator muscle were considered. The methods applied were min-max, 1st-99th percentiles and 3-Sigma normalization, z-score standardization, mean centering, histogram normalization, Nyul-Udupa and ComBat harmonization. The Mann-Whitney U-test was applied to assess features repeatability between scanners, by comparing the feature values obtained for each normalization method, including the case in which no normalization was applied. Most image normalization methods allowed to reduce the overall variability in terms of intensity distributions, while worsening or showing unpredictable results in terms of feature robustness, except for the *z-score*, which provided a slight improvement by increasing the number of statistically similar features from 9/93 to 10/93. Conversely, feature normalization methods positively reduced the overall variability across the scanners, in particular, 3*sigma*, *z_score* and *ComBat* that increased the number of similar features (79/93). According to our results, it emerged that none of the image normalization methods was able to strongly increase the number of statistically similar features.

1. Introduction

In the past decades, the field of medical image analysis has grown exponentially and is gaining more and more importance in disease diagnosis and patient care (Gillies *et al* 2016, Lu *et al* 2019). A great role is played by radiomics, that involves the extraction and analysis of quantitative image features describing the patterns of pixel intensity variations within an image, through the use of a series of mathematical algorithms (Haralick *et al* 1973, Lambin *et al* 2012, 2017). Several efforts have been made towards the development of radiomics-based systems for supporting cancer diagnosis, treatment planning, and response assessment, that could be applied to routinely acquired medical images (Stanzione *et al* 2022). Despite the spread of studies involving this kind of systems, one of the main limitations of their introduction in clinical practice is the poor reproducibility, generalizability, and robustness of radiomics features, especially in multi-center setting (Fusco *et al* 2022, Stamoulou *et al* 2022). The main reason is due to the unavoidable image variability

caused by both biological and non-biological factors, e.g., scanners, acquisition protocols, pre-processing software, etc, which may lead important bias.

Currently, several efforts have been made for solving this issue, and two main approaches have been developed: *image normalization* and *feature normalization* (Alrahawy *et al* 2022, Stamoulou *et al* 2022, Stanzione *et al* 2022). The first approach consists in applying a normalization method to each image or volume to bring pixel intensities into a common range or distribution before extracting features. Conversely, during feature normalization the values of features are rescaled to obtain ranges or distributions similar for all centers. With respect to the first approach, in this case a subgroup of training samples for each center is needed to estimate the normalization parameters. However, in the last years few studies proposed standardized pipelines or guidelines to harmonize differences among patients and scanners for both positron emission tomography (PET) and computed tomography (CT) (Traverso *et al* 2018, Ly *et al* 2019, Da-Ano *et al* 2020, Kociołek *et al* 2020). Conversely, no standardized procedures have been proposed for magnetic resonance imaging (MRI) (Stamoulou *et al* 2022). Therefore, further researches are needed to define possible procedures for MRI, in which problems related to feature robustness are accentuated since intensities are non-standardized and highly dependent on manufacturer and acquisition protocol parameters (Scalco and Rizzo 2017).

Some studies have been done in this direction, analyzing the effects of different normalizations on the radiomics features extracted on MRI: some of them consider databases related to the brain (Reinhold *et al* 2019, Carré *et al* 2020), other consider MRI sequences related to different body parts (thorax and abdomen) acquired by the same scanner in different times (before and after radiotherapy) (Chatterjee *et al* 2019, Schwier *et al* 2019, Upadhaya *et al* 2019, Isaksson *et al* 2020, Scalco *et al* 2020, Mchugh *et al* 2021, Campello *et al* 2022, Granzier *et al* 2022), and others on phantom databases (Buch *et al* 2018, Rai *et al* 2020). However, most of their insights might be biased by the choices made in terms of model selected and related parameters. Furthermore, they do not truly assess how to reduce the variability among different scanners, since most of their databases are mono-scanner.

To the best of our knowledge, only two studies focused on the robustness of the features across different centers in a generalizable manner (Crombé *et al* 2020, Li *et al* 2021), evaluating the differences between normalization methods on multi-center brain MRI datasets. Unfortunately, there are not similar studies related to the abdominal area in a multi-center setting.

The aim of this study is to evaluate how different image and features normalization methods impact the robustness of radiomics features acquired using a multicenter and multi-scanner abdominal MRI dataset.

2. Material and methods

2.1. MRI dataset

Clinical rectal scans were retrospectively collected from patients with histologically confirmed stage II/III Locally Advanced Rectal Cancer (LARC) in three different hospitals: Candiolo Cancer Institute, FPO-IRCCS of Candiolo (Italy) (Center A); Mauriziano hospital of Turin (Italy) (Center B); Molinette hospital A.O.U. Città della Salute e della Scienza of Turin (Italy) (Center C). The following MRI scanners were adopted:

- Scanner A.1: 1.5 T GE scanner using an 8-channel phased-array surface coil (HDx Signa Excite, GE HealthCare, Milwaukee, WI, USA) in center A;
- Scanner A.2: 1.5 T GE scanner using a 32-channel phased-array surface coil (Optima MR450w, GE HealthCare, Milwaukee, WI, USA) in center A;
- Scanner B: 1.5 T Philips scanner using a 32-channel body phased-array coil (Ingenia, Philips Medical Systems, Eindhoven, The Netherlands) in center B;
- Scanner C: 1.5 T Philips scanner using a 32-channel body phased-array coil (Achieva, version 2.6, Philips Medical Systems, Eindhoven, The Netherlands) in center C.

All standard sequences were collected, according to MRI guidelines for reporting rectal cancer staging (Beets-Tan *et al* 2018), however in this study we considered the fast spin-echo T2 weighted (T2w) sequence acquired on the axial plane perpendicular to the longest tumor diameter. Parameters of T2w sequence are reported in table 1 for each institution.

The study was approved by the institutional review boards in each institution, with a waiver for requirement of informed consent as de-identified patient data were utilized.

2.2. Normalization approaches

In this study we compared two different normalization approaches, one based on image normalization and the other based on features normalization. For each approach, the most widely used methods were applied to our sequences. The study flowchart is depicted in figure 1 and each step is detailed below for both approaches.

2.3. Image normalization approach

2.3.1. Image crop

First, all slices of each patient were centered cropped to obtain the same Field Of View (FOV), i.e. $220 \times 220 \times 100$ mm^3 that is the smallest among the four scanners and includes the abdominal area. In this way, we obtained images that include approximately the same anatomical structures to ensure that, when we will apply image normalization methods based on image intensities, results were not biased by the inclusion of different organs that have different signal intensities. An example of the result of the crop is shown in figure 2.

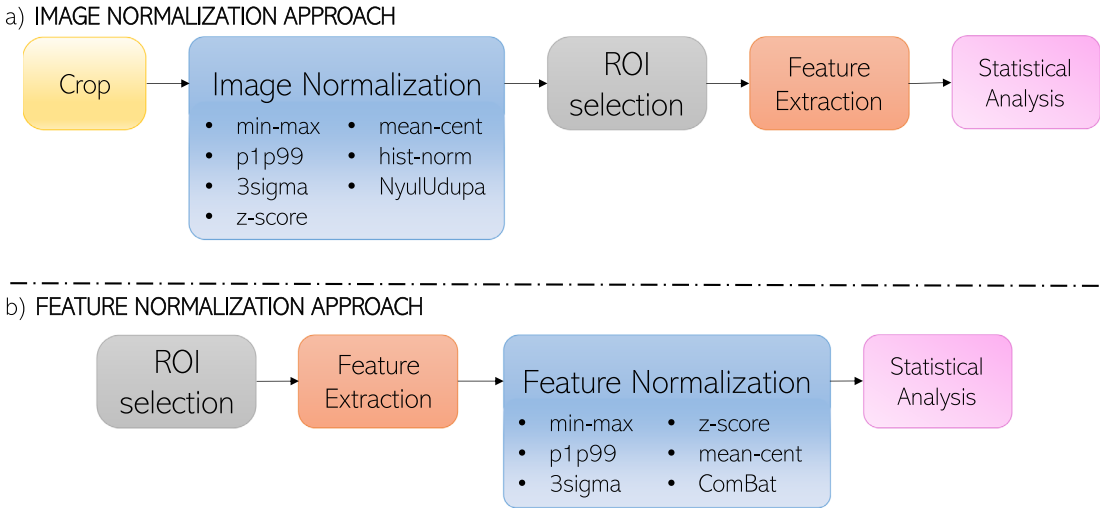

Figure 1. Flow-chart of the approaches followed in the study: a) image normalization steps, b) feature normalization steps.

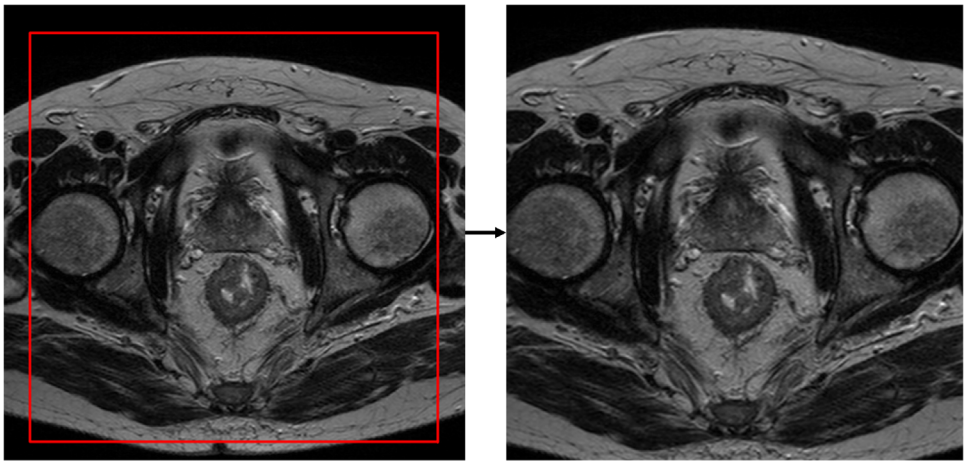

Figure 2. Example of the preprocessing applied to T2-weighted rectal MRI. On the left there is the original image from center C (FOV = 250 mm × 250 mm), and on the right the cropped one (FOV = 220 mm × 220 mm).

Table 1. Characteristics of the T2w sequences of each scanner.

Parameters	Scanner A.1	A.2	B	C
TR/TE	3740/110 ms	7660/110 ms	3231/90 ms	5085/100 ms
Acquisition matrix	416 × 224	416 × 224	320 × 311	512 × 512
Slice thickness	4.4 mm	4 mm	3.5 mm	3 mm
Slice spacing	4.4 mm	4 mm	3.85 mm	3.3 mm
Pixel size	0.43 × 0.43 mm^2	0.43 × 0.43 mm^2	0.47 × 0.47 mm^2	0.49 × 0.49 mm^2
Pixel bandwidth	162.773 Hz	244.14062 Hz	328 Hz	126 Hz
FOV	220 mm × 220 mm	220 mm × 220 mm	240 mm × 240 mm	250 mm × 250 mm
NEX	2	3	1	2
Flip angle	90°	90°	90°	90°
T2w dimension	512 × 512	512 × 512	512 × 512	512 × 512

T2w = T2 weighted, TR = Repetition time, TE = Echo Time, FOV = Field Of View.

2.3.2. *Image normalization*

Seven image normalization methods were applied to each cropped volume (Stamoulou *et al* 2022, Stanzione *et al* 2022):

- Min-Max (*min_max*) that rescales the intensities into the range [0, 1], using equation (1):

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where z_i is the i-th voxel's normalized intensity, x_i is the i-th voxel's original intensity, x_{min} and x_{max} are respectively the minimum and the maximum intensity within the volume.

- p1-p99 (p1p99) that rescales the intensity range using the 1st and 99th percentile of the intensity distribution, using equation (2):

$$z_i = \frac{x_i - x_{p1}}{x_{p99} - x_{p1}} \quad (2)$$

where z_i is the i-th voxel's normalized intensity, x_i is the i-th voxel's original intensity, x_{p1} and x_{p99} are respectively the 1st and 99th percentiles of the intensities within the volume.

- 3-Sigma (3sigma) that rescales the intensity range using the mean and the standard deviation multiplied by three, using equation (3):

$$z_i = \frac{x_i - (\bar{x} - 3\sigma)}{(\bar{x} + 3\sigma) - (\bar{x} - 3\sigma)} \quad (3)$$

where z_i is the i-th voxel's normalized intensity, x_i is the i-th voxel's original intensity, $\bar{x}$ and σ are respectively the mean and the standard deviation of the intensities within the volume.

- Z-Score (z_score) that standardizes the intensities distribution to have zero-mean and unit-variance, using equation (4):

$$z_i = \frac{x_i - \bar{x}}{\sigma} \quad (4)$$

where z_i is the i-th voxel's normalized intensity, x_i is the i-th voxel's original intensity, $\bar{x}$ and σ are respectively the mean and the standard deviation of the intensity distribution of the volume.

- Mean Centering (mean_cent) that standardizes the intensity values by subtracting the mean value of the volume intensities, using equation (5):

$$z_i = x_i - \bar{x} \quad (5)$$

where z_i is the i-th voxel's normalized intensity, x_i is the i-th voxel's original intensity and $\bar{x}$ is the mean of the intensity distribution of the volume.

- Histogram Normalization (hist_norm) that reshape the volumes histogram using the histogram of a specific volume as reference. This process consists of three steps:

1. Compute the histogram of the volume (H_{img}) and the reference (H_{Ref}).
2. Compute the discrete cumulative distribution functions (CDFs) of both histograms, CDF_{img} and CDF_{ref}.
3. Compute a mapping that transforms the intensity distribution of the volume so that it matches the intensity distribution of the reference. This is obtained by minimizing the following equation (equation (6)):

$$M(x_{img}) = \text{argmin} \left| CDF_{img}(x_{img}) - CDF_{ref}(x_{ref}) \right| \; \forall \; x_{img} \in \text{Volume} \quad (6)$$

In this study, we evaluated the effects of histogram normalization changing the reference sequence. Therefore, we repeated this analysis 4 times using iteratively one sequence randomly selected from each scanner as reference.

- Nyul-Udupa Normalization (NyulUdupa), that was first presented by Nyul *et al* (Nyú and Udupa 1999) and it is composed of two steps:

1. *Training,* where the parameters (landmarks) of a 'standard' histogram are estimated from the training volumes. In case of a bimodal distribution, the landmarks are: the minimum and maximum intensities, the 1st and the 99th percentile of the intensities, and the second mode of the histogram. In this study, we evaluated the effects of this image normalization changing the training set. Therefore, we repeated this analysis 4 times, using all sequences from each scanner for the training.
2. *Transformation,* where volume v is transformed so that its histogram parameters match those of the 'standard' one. This step is performed by a non-linear intensity transformation in which two separate mappings are applied according to the voxel's original intensity x_i (equation (7)):

$$z_i = \begin{cases} \left[\mu_s + (x_i - \mu_v) \frac{s_1 - \mu_s}{p_{1v} - \mu_v} \right], & \text{if } m_{1v} \leqslant x_i \leqslant \mu_v \\ \left[\mu_s + (x_i - \mu_v) \frac{s_2 - \mu_s}{p_{2v} - \mu_v} \right], & \text{if } \mu_v \leqslant x_i \leqslant m_{2v} \end{cases} \quad (7)$$

where z_i is the voxel's normalized intensity, m_{1v} and m_{2v} are the minimum and maximum intensities on the volume v, p_{1v} and p_{2v} are the 1st and the 99th percentile of the intensities on the volume v, s_1 and s_2 are the minimum and the maximum intensities of the standard histogram, and μ_v and μ_s are the second mode of the histogram of the original volume v and that of the standard histogram respectively.

2.3.3. Regions of Interest (ROIs) selection

For each patient, we manually selected six different regions of interest (ROIs) ($10 \times 10 \times 3$ voxels) belonging to the obturator muscle using ITK-SNAP (http://www.itksnap.org/pmwiki/pmwiki.php) (Yushkevich *et al* 2006) on different slices (figure 3). These ROIs were identified on the original sequences and identically

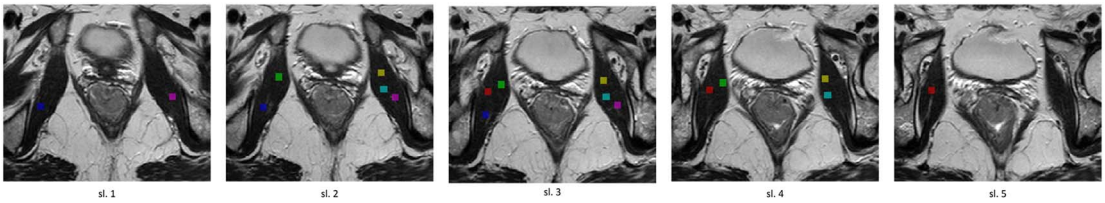

Figure 3. Examples of labeled ROIs on T2w sequence of patient id. 2032. Each ROI was labelled with a number: 1 - dark blue, 2 – pink, 3 – green, 4 – yellow, 5 – light blue, 6 – red.

selected on each normalized sequence. We choose to take into account the obturator muscle since it is a clearly identifiable homogeneous healthy structure, with similar characteristics between men and women of different ages. Moreover, since it is not adjacent to the tumor, its signal intensities are not affected by the presence and characteristics of the tumor (Wang *et al* 2008, Giannini *et al* 2016). Due to these stable properties, it has been used as reference tissue to normalize intensities on T2w images (Engelhard *et al* 2000, Dikaios *et al* 2015, Giannini *et al* 2016).

2.3.4. Feature extraction

Ninety-three features were computed on each ROI for all normalized sequences and for the original sequences for comparison. Feature extraction was performed using the IBSI compliant open-source platform PyRadiomics (Breiding 2014), implemented in Python 3.7. In particular, the following classes of first and second order features were considered:

- **First Order Statistics** (18 features): they describe the distribution of voxel intensities within the ROI.
- **Gray Level Co-occurence Matrix**– GLCM (24 features): it describes the second-order joint probability function of a ROI definied as P(i,j|δ,θ), where the (i,j)th element of this matrix represents the number of times the combination of levels i and j occur in two pixels in the image, that are separated by a distance of distance δ along the angle θ.
- **Gray Level Dependence Matrix**—GLDM (14 features): it quantifies the gray level dependencies in a ROI. A gray level dependency is defined as the number of connected voxels within a given distance δ that are dependent on the center voxel.
- **Gray Level Run Length Matrix**—GLRLM (16 features): it quantifies the gray level runs, which are defined as the length of consecutive voxels that have the same gray level value.
- **Gray Level Size Zone Matrix**– GLSZM (16 features): it quantifies the gray level zones in a ROI. A gray level zone is defined as the number of connected voxels that share the same gray level intensity.
- **Neighbouring Gray Tone Difference Matrix**—NGTDM (5 features):it quantifies the difference between a gray value and the average gray value of its neighbors within a given distance δ.

The discretization was performed considering a fixed bin count, thus allowing a direct comparison of the feature values across multiple analyzed ROIs (Zwanenburg *et al* 2016). In particular, we used 32 bins due to the expected ranges of the pixel intensity values. The features extraction was performed in 2.5D to avoid using interpolated isotropic voxels (Zwanenburg A, Leger S, Vallières M 2016). Finally, we used $\delta = 1$ for all matrices and $\alpha = 0$ for GLDM, to consider the nearest neighborhoods, given the small ROIs sizes.

2.4. Feature normalization approach

2.4.1. ROIs selection

Starting from the original sequences, for each patient we selected exactly the six ROIs used for the image normalization approach.

2.4.2. Feature extraction

For each of the six ROIs, the 93 features already used for the image normalization approach were computed.

2.4.3. Feature normalization

Six features normalization methods were applied on the features extracted from the original sequences. As suggested by (Chatterjee *et al* 2019), all normalizations were applied to each center separately. To avoid overfitting, the normalization parameters (minimum, maximum, mean value, etc) were extracted from a subgroup of 60% of sequences for each center and applied to the entire set of features from the same center.

The following features normalization methods were compared:

- Min-Max (*min_max*) that rescales the feature into the range [0, 1], using equation (1), where z_i is the i-th feature's normalized value, x_i is the i-th feature's original value, x_{min} and x_{max} are respectively the minimum and the maximum feature value within the scanner subgroup.

- p1-p99 (*p1p99*) that rescales the feature range using the 1st and 99th percentile of the feature value distribution, using equation (2), where z_i is the i-th feature's normalized value, x_i is the i-th feature's original value, x_{p1} and x_{p99} are respectively the 1st and 99th percentiles within the scanner subgroup.
- 3-Sigma (*3sigma*) that rescales the feature range using the mean and the standard deviation multiplied by three, using equation (3), where z_i is the i-th feature's normalized values, x_i is the i-th feature's original value, $\bar{x}$ and σ are respectively the mean and the standard deviation of the values within the scanner subgroup.
- Z-Score (*z_score*) that standardizes the feature values distribution to have zero-mean and unit-variance, using equation (4), where z_i is the i-th feature's normalized value, x_i is the i-th feature's original value, $\bar{x}$ and σ are respectively the mean and the standard deviation of the values within the scanner subgroup.
- Mean Centering (*mean_cent*) that standardizes the intensity values by subtracting the mean value of the scanner subgroup ($\bar{x}$), using equation (5), where z_i is the i-th feature's normalized value, x_i is the i-th feature's original value.
- ComBat (*ComBat*), an harmonization method, originally developed for genomics, which corrects the differences in radiomics features due to the so-called *batch effects*, i.e., different scanners and acquisition protocols (Horng *et al* 2022, Orlhac *et al* 2021). It is a data-driven realignment transformation following equation (8):

$$y_{ij}^{\mathrm{ComBat}} = \frac{y_{ij} - \hat{\alpha} - \widehat{\gamma_i}}{\widehat{\delta_i}} + \hat{\alpha} \qquad (8)$$

where $\hat{\alpha}$, $\widehat{\gamma_i}$ and $\widehat{\delta_i}$ are estimators of the feature average value, additive batch effect and multiplicative batch effect, respectively, and y_{ij}^{ComBat} is the transformed y_{ij} j-th feature devoid of the i-th scanner effect (Orlhac *et al* 2021). For the evaluation of the estimators, all subgroups from each center were pooled together. For the ComBat application, we used the *neuroComBat* Python package(Fortin *et al* 2018).

2.5. Statistical analysis

In order to assess whether the features from different scanners have the same distributions, the Mann-Whitney U-test (Nachar 2008) was applied to compare the values obtained in the 6 ROIs for all patients using two different scanners. Therefore, since we considered 4 different scanners, a total of 6 pairwise comparisons were obtained. The Bonferroni correction was applied to take into account the presence of multiple ROIs belonging to the same patient (Curtin and Schulz 1998). A p-value lower than 0.05 was considered as significant. This procedure was repeated for each normalization method, and for the case in which no normalization was applied (*no_norm*) for comparison.

3. Results

88 patients (56 men and 32 women) were retrospectively collected, having an average age of 64 years (range 37–83): 24 patients from scanner A.1, 15 from A.2, 22 from B and 27 from C.

3.1. Image normalization

Figure 4 shows the heatmap representing the percentage of not statistically different features before (*no_norm*) and after applying image normalization. Without applying image normalization, first order features are the least similar, which is consistent with literature, however also NGTDM features showed the lowest percentage of similar features. This behavior can be due to the sample size of this group of features (N = 5) that could affect the values of the percentages. However, when using a similar protocol on different scanners of the same manufacturer (A.1 versus A.2), all second order features are stable also without any normalization. Despite we have applied different normalization methods based on either image intensities rescaling or standardization or histogram remapping, we observed that any of them was able to strongly increase the number of similar features, as we expected at least on first order features. Considering the total number of similar features across all comparisons, we observed different and interesting behaviors: the highest increase of similar features, compared to *no_norm*, was obtained with *hist_norm* but only when center B was used as reference (194 versus 162 similar features across all comparisons). Conversely, the use of other centers as reference has unpredictably caused either an increase (178 with center A.2 taken as reference) or decrease (144 and 147 with center C and A.1 as reference, respectively) in similar features. However, if we consider each pairwise comparison of centers (figure 5, panel B), we can see that the increase in similar features obtained by *hist_normB* is caused by a substantial increase between centers A2 versus C, A2 versus B and B versus C, offset by a decrease between A1vsA2 and A1vsB. On the other side, *3_sigma* and *Z_score* obtained the second and the third highest number of similar features (176 and 175 respectively), without being dependent on the choice of a reference. The worst results were obtained with the *NyulUdupa* producing from 159 to 165 similar features across all comparisons, depending on the reference used.

Figure 5 shows the total number of not statistically different features grouped by normalization method (panel A) and comparison among scanners (panel B). Focusing on the results obtained with *NyulUdupa*, it emerged that the reference sequences does not affect the number of similar features, however none of

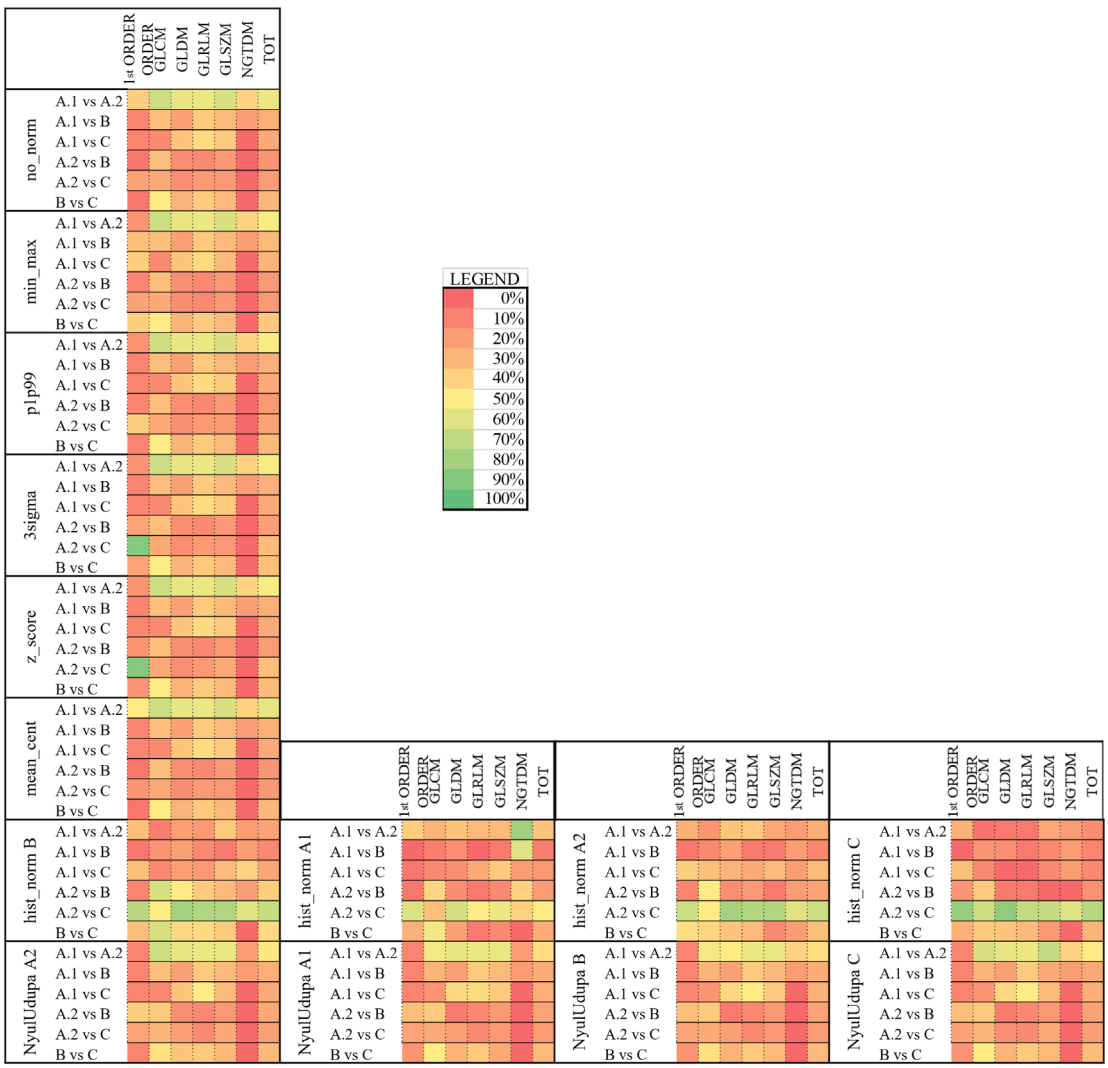

Figure 4. Heatmap of the percentage of not statistically different features resulting from the Mann-Whitney test, related to the image normalization methods.

reference allowed to obtain an increase of similar features.

Figure 5 panel B confirms that a similar protocol allows to obtain the best results (A.1 versus A.2), except for *hist_norm*, even if no methods allow to increase the number of similar features obtained without any image normalization. Vice versa, all comparisons with center B and C, except for *hist_norm*, produce the worst results, also when images are acquired with different protocols on scanners produced by the same manufacturer (B versus C). Analyzing the acquisition parameters reported in table 1, it emerges that images from center B are characterized by lower signal to noise ratio (SNR) due to, for example, lower number of excitations (NEX) and higher pixel bandwidth that may have affected the quality of the images.

In Supplementary figure 1 we reported the number of times for which each feature is not statistically different among the four scanners. Interestingly, features resulting statistically similar for at least five comparisons show the same behavior with and without normalizations, except *hist_norm*. In particular, these features are: Skewness (First Order); ClusterShade, JointAverage, and SumAverage (GLCM); HighGrayLevelEmphasis (GLDM); HighGrayLevelRunEmphasis and ShortRunHighGrayLevelEmphasis (GLRLM); GrayLevelVariance and HighGrayLevelZoneEmphasis (GLSZM). Of note, among all normalizations, z-score slightly increases the number of similar features with respect to the no-norm case (10/93 versus 9/93), while *hist_norm* heavily decreases it (4/93 versus 9/93). This is in line with we previously said about *hist_norm* because we showed that this method obtained different results depending on the pairwise comparisons, causing a decrease of similar features for at least 5 combinations. Even if the *NyulUdupa* method affects the histogram distribution of the images like *hist_norm*, the results of the comparisons are similar to the *no-norm* case, since it maintains the meaning of tissue intensities (Nyú and Udupa 1999).

The impact of each image normalization method on the image histograms is shown in figure 6. Among

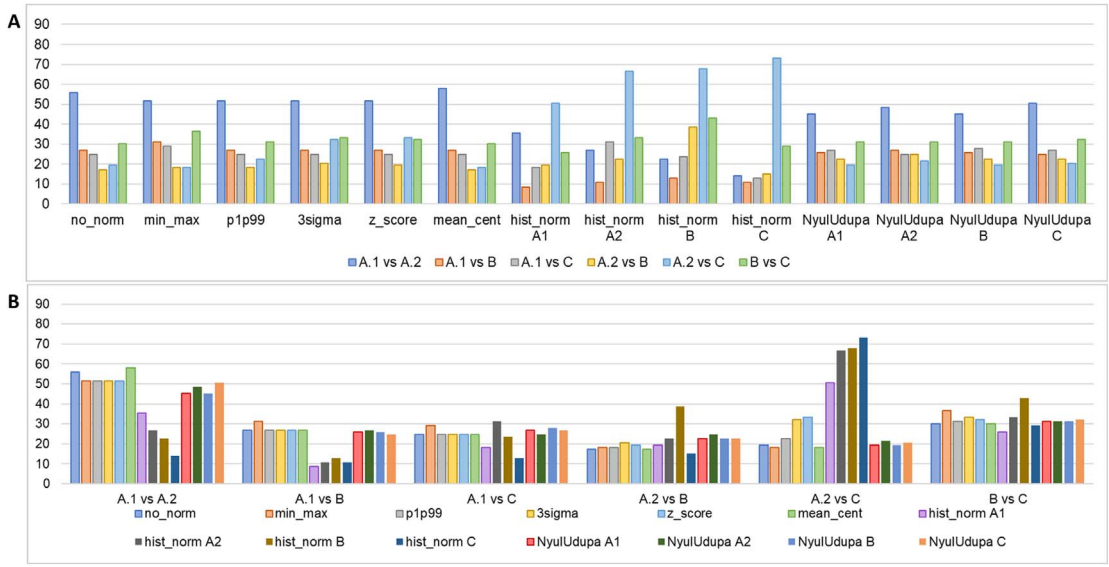

Figure 5. Total number of not statistically different features grouped by normalization method (panel A) and comparison among scanners (panel B).

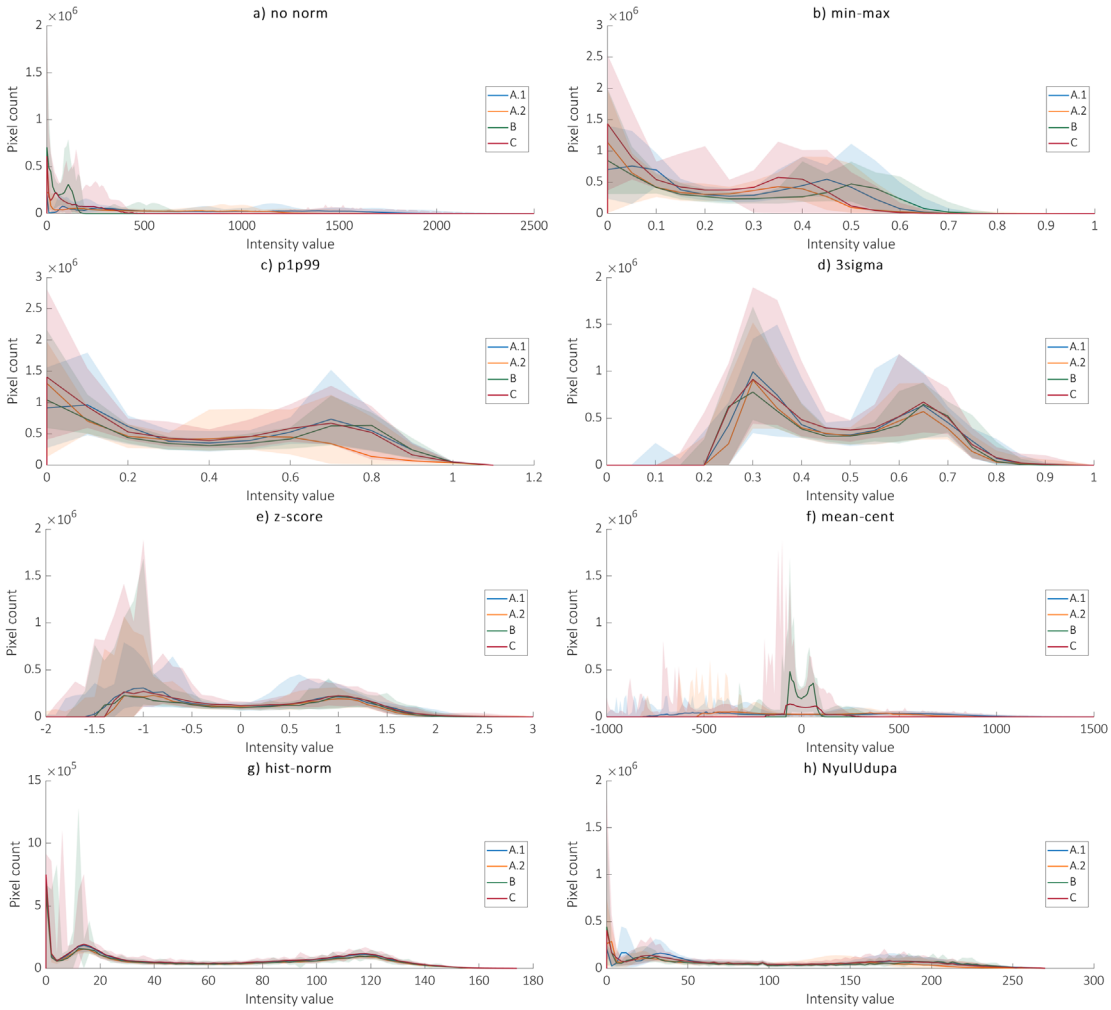

Figure 6. Image histograms of abdominal MRI, acquired with 4 different scanners. The subfigures show the image histograms of the MRI images without apply any normalization method (a), and after the min-max (b), the p1p99 (c), the 3sigma (d), the z-score (e), the mean centering (f), the histogram (g), and Nyul-Udupa (h) normalization methods. The solid lines represent the median histograms across all patients acquired with a given scanner, whereas the colored areas represent the corresponding max-min ranges.

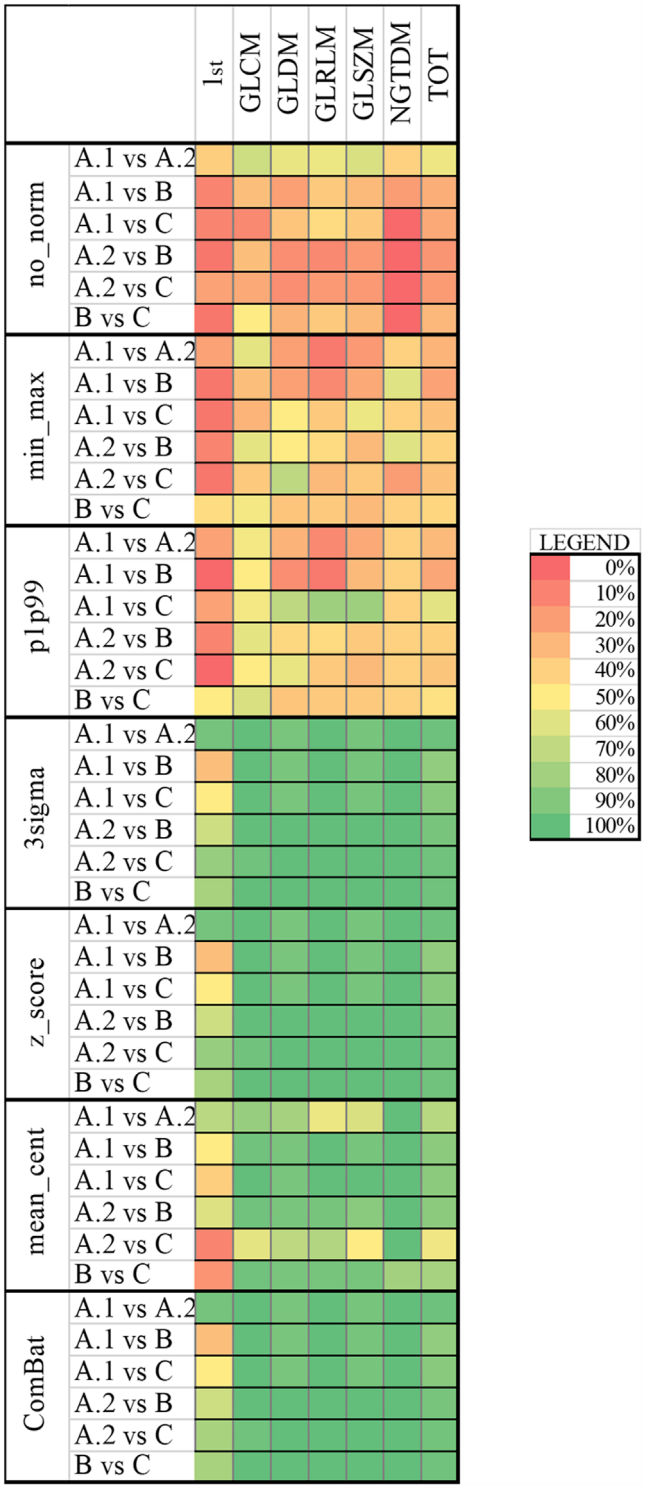

Figure 7. Heatmap of the percentage of not statistically different features resulting from the Mann-Whitney test, related to the feature-based normalization methods.

methods which do not use a histogram reference, both *3sigma* and *p1p99* obtained better results, since they allow both to obtain more similar median histograms and to reduce the variability among scanners (see the colored areas in figure 6). Considering the other methods based on histogram references, i.e., *hist-norm* and *NyulUdupa*, they both allow to obtain very similar median histograms, although there are still some variabilities due to borderline cases in each scanner.

3.2. Feature normalization

Figure 7 shows the heatmap based on the percentage of not statistically different features before (*no_norm*) and after applying features normalization. Since for the *ComBat* method the average across all centers is taken as the reference batch (Fortin *et al* 2018), results obtained by applying different harmonization references are identical. Therefore, we will report only the results related to the ComBat_A1 (*ComBat*). All

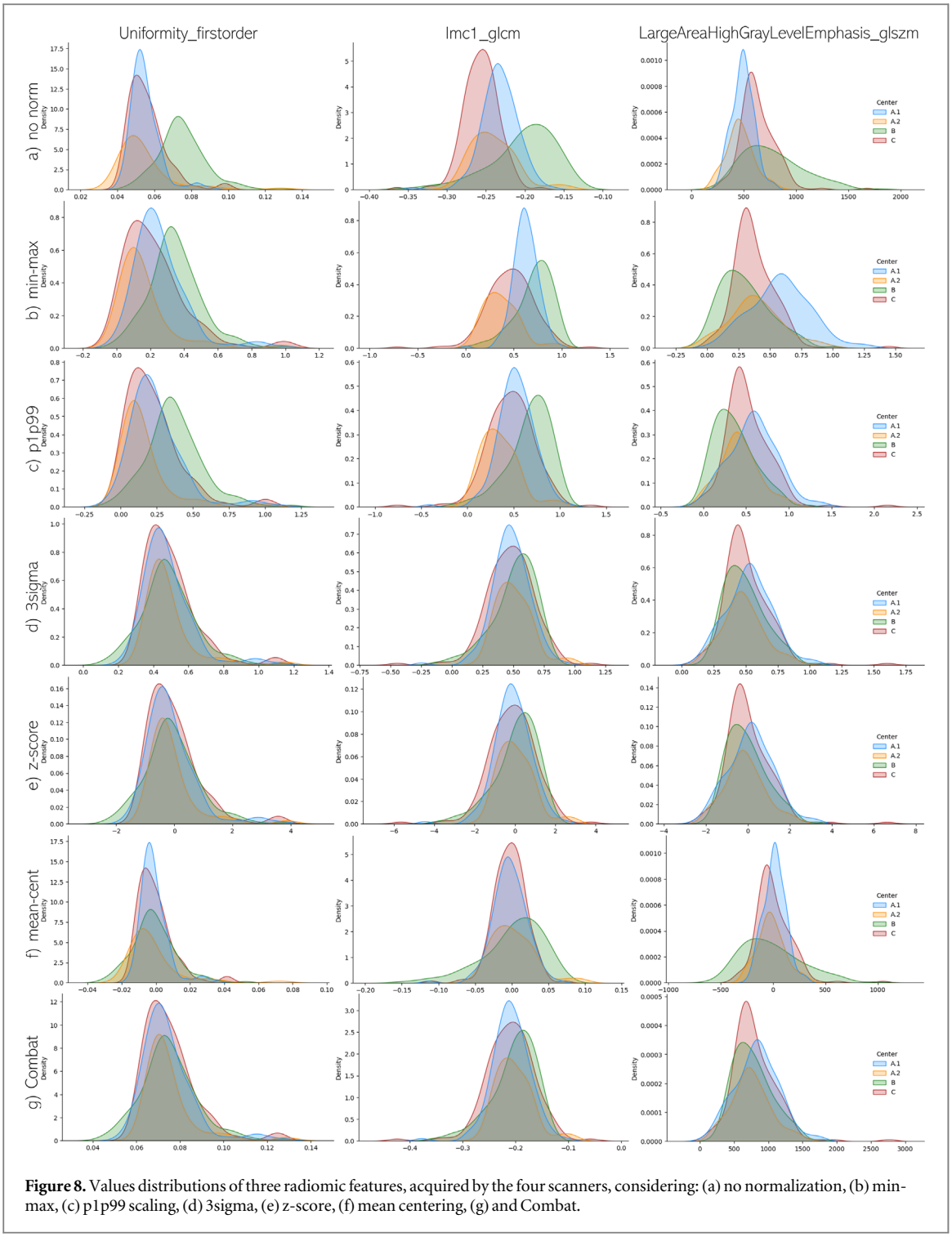

Figure 8. Values distributions of three radiomic features, acquired by the four scanners, considering: (a) no normalization, (b) min-max, (c) p1p99 scaling, (d) 3sigma, (e) z-score, (f) mean centering, (g) and Combat.

methods produce an increase of the number of similar features with respect to the *no_norm* condition, for all comparisons. In particular, *3sigma*, *z_score* and *ComBat* provided the highest number of similar features across all comparisons (516, 516, and 515 respectively). The *mean_cent* provided good results for all second order features, except for the comparison between A.1 and A.2 and between A.2 and C. Also in this case, the first order features are the most difficult to make similar among different centers, in fact they presented the lowest number of similar distributions. However, when using a similar protocol on different scanners of the same manufacturer (A.1 versus A.2), also the first order features are positively affected by feature normalization methods based on *3sigma*, *z_score* and *ComBat* since they heavily impact the feature distribution rescaling mean and standard deviation and reducing differences among scanners. In figure 8 we show the distributions of three radiomics features (Uniformity_firstorder, Imc1_glcm, and LargeAreHighGrayLevelEmphasis_glszm) obtained with and without feature normalization. In particular, it is possible to observe how the distributions are affected by each method: *min-max* and *p1p99* only

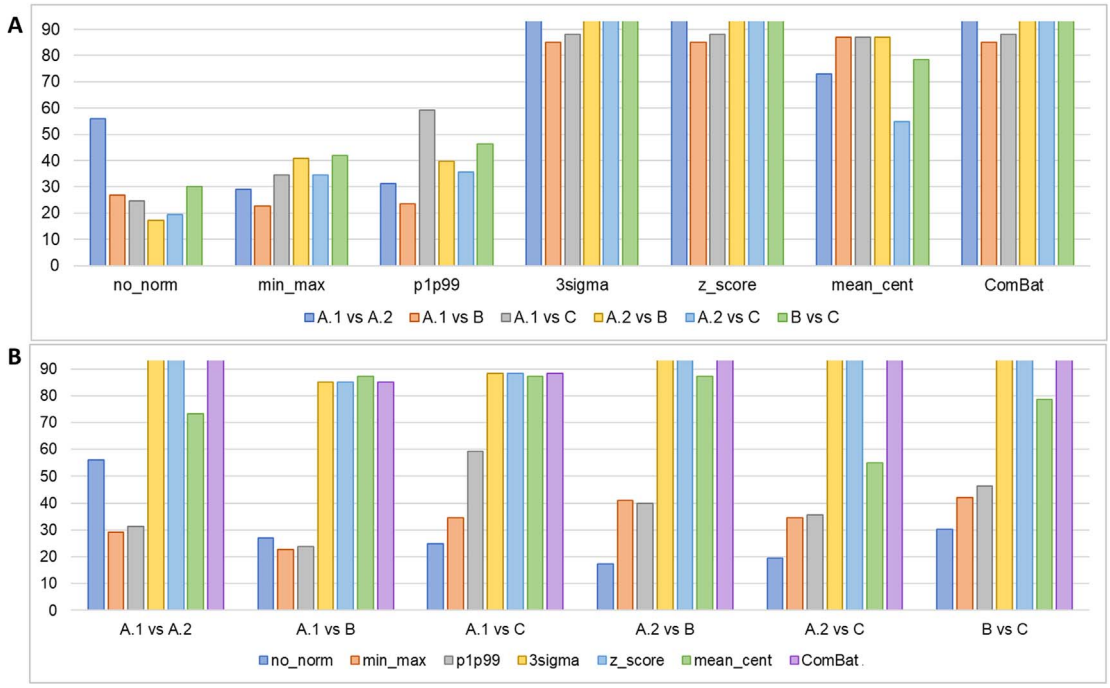

Figure 9. Total number of not statistically different features grouped by normalization method (panel A) and comparison among scanners (panel B).

impact the feature range values, while *3sigma, z-score, mean-cent* and *ComBat* positively impact all distributions, reducing the differences among scanners.

Figure 9 shows the total number of not statistically different features grouped by normalization method (panel A) and comparison among scanners (panel B). From both panels it emerged that *min_max* and *p1p99* methods provided a slightly improvement for most comparisons, except for A.1 versus A.2 and A.1 and B, while *mean_cent* provided poorer results with respect to *3sigma*, *z_score* and *ComBat*, due to the fact that this normalization simply rescales the mean, not affecting the shape of the distribution.

In Supplementary figure 2 we reported the number of times for which each feature is not statistically different among the four scanners. It is possible to notice that the number of similar features in at least 5 comparisons is increased by *3sigma*, *z-score* and *ComBat*, with respect to the *no-norm* case (79/93 versus 9/93). In particular, they reduced the differences across the scanners for almost all texture features, except for LargeDependenceLowGrayLevelEmphasis (GLDM) and LargeAreaLowGrayLevelEmphasis (GLSZM). Regarding the first order group, only six features out of 18 resulted similar in at least 5 comparisons, i.e., entropy, kurtosis, maximum, range, skewness and uniformity.

4. Discussion

In this study we assessed the effect of two normalization approaches and related methods on a set of first and second order radiomics features extracted from an abdominal multicenter MRI dataset. In particular, we evaluated if the internal variability of a multi-vendor and multicenter database could be reduced by applying either image or feature normalizations.

According to our results, it emerged that none of the image normalization methods was able to strongly increase the number of statistically similar features. The best improvement was obtained by the *z-score* that brought the number of features statistically similar for at least five comparisons from 9/93 to 10/93. Indeed, the median number of statistically similar features among all normalization methods was 9/93 (IQR: 5–9). In addition, some normalizations methods have slightly worsened the results in specific comparisons (e.g., *hist_norm* for the comparison A2 versus C). The reason of this behavior is not fully explainable, and it might depend also on the initial differences between the histogram of the original sequence and that used as reference. Even if the features distributions were not heavily affected by the image normalizations, we observed a slight improvement in similarity on histograms when using *3sigma*, and *p1-p99*. This insight could be helpful in particular during the development of deep learning models, which learn directly from images. However, since the tissues included in the sequences heavily influence the normalization parameters in case of image normalization, we strongly recommend ensuring similar FOVs for all analyzed images, as presented in the paper.

Conversely, normalizing features rather than images allows to reduce their variability, providing

good reproducibility among different centers and scanners. The best results were obtained with *3sigma*, *z_score* and *ComBat* rather than with methods based on rescaling the values, i.e., min-max. These findings are in accordance with those by (Chatterjee *et al* 2019), that compared the effects of feature rescaling and standardizing proving the improvement of the predictive radiomics model given by standardizing the features separately for each independent set. Moreover, these results are consistent with theory, since these methods are based on parameters (i.e., mean, standard deviation, the estimators of the *ComBat* algorithm) describing the overall data distribution on the training set rather than on the extreme values such as *min-max* and *p1-p90*.

The highest increase of similar features after normalization was reached by first-order's group with some features normalization methods (*3sigma*, *zscore*, *mean_cent* and *ComBat*). Conversely, image normalization was not able to increase reproducibility across centers neither for first order features. These results brought an advance in knowledge, since theory suggests that image normalization should impact first order features, however to the best of our knowledge no previous studies proved this point on a clinical multicenter dataset.

Previously, (Rai *et al* 2020) and (Buch *et al* 2018) assessed, using phantoms, the reproducibility of first and second order features as MRI scanners and acquisition parameters change. In particular, (Rai *et al* 2020) obtained that, using a controlled protocol and a phantom developed for that purpose, first order features were the most robust among scanners. This seems to be in contrast with our results, but if we compare A1 versus A2 scanners (two scanners in the same center with very similar acquisition protocols) we obtained an increase of similar first order features with some image and features normalization methods (i.e., *mean_centering, hist_norm, 3sigma*, *z_score* and *ComBat*). Focusing on the texture features, they found substantial changes in the robust features subset due to use of different phantoms, suggesting that it is not possible to develop a common normalization approach for all MRI-based radiomics algorithms if they will be applied on different organs. Similarly, (Buch *et al* 2018) evaluated differences in radiomics features using the same phantom and changing scanner platform and scan parameters such as magnet strength, flip angle, and NEX, which is a more representative situation of the available clinical datasets. The most substantial changes in the features were encountered with differences in MRI scanner manufacturer and NEX and these results are consistent with our findings. In our study, the strongest differences were found between A.1/A.2 and B, in which both parameters were different, and between B and C scanners that share the same manufacturer but using different NEX.

Other similar studies were conducted assessing the impact of normalization on radiomics features, yielding almost similar results. In particular, (Scalco *et al* 2020) carried the analysis on abdominal MRIs acquired on two different times (before and after radiotherapy) with the same scanner whereas (Granzier *et al* 2022) performed a similar analysis on a multi-scanner breast MRI dataset. Both studies concluded that *z-score* image normalization method provides the most reproducible features on MRI sequences. Similarly to us, (Li *et al* 2021) compared six image normalization methods and the *ComBat* on brain radiomics features, assessing that the latter positively affected the feature robustness across different acquisition scanners. Moreover, they underlined that, even if the feature robustness was not significantly affected by the image normalizations, they allow to obtain comparable brain MRI images by bringing the image intensities into a common scale, thus overcoming the non-standardize and interpretable MRI intensities.

A more comprehensive comparison with literature is difficult because of the limited number of studies evaluating the impact of image and feature normalization on abdominal MRI and the lack of multicenter analysis. Moreover, most previously mentioned studies used either the Intraclass Correlation Coefficient (ICC) or the Student t-test to assess features reproducibility. However, the ICC, similarly to the Student t-test, is subject to statistical assumptions such as normality and/or stable variance, which are rarely considered in health applications (Bobak *et al* 2018). In addition, ICC assesses the reliability of a measure, defined as the extent to which measurements can be replicated (Koo and Li 2016). This means that exactly the same element or patient must be evaluated by different raters or using different devices. In this study we compared the feature values calculated on different sets of patients (those acquired by each center), therefore ICC results not appropriate for this purpose. Only (Li *et al* 2021) used the Wilcoxon test, similarly to us, to measure the similarity between the distributions of features extracted by the healthy white matter regions, which are not affected by gender, age, and the tumor's presence and characteristics. In literature, researchers almost unanimously agree on the importance of a normalization step for radiomics-based system development, above all in multicenter studies. However, no common and clear indications can be found about what kind of normalization method is better for a given application. This work aimed to fill this gap, giving evidence of what can happen in clinical situations, in order to make more conscious choices. In fact, compared to previous studies, results of our analysis better reflect what could happen when developing radiomics-based systems for the clinical practice, since multicenter images acquired using different protocols should be managed. In this context, we demonstrated that image normalization methods do not allow to overcome the issue of features

reproducibility across different centers. Conversely, using a proper features normalization method, i.e., *3sigma*, *zscore*, *mean_cent* and *ComBat* can strongly affect the number of similar features. This behavior might positively impact the transability of these systems in clinical practice, allowing the development of more robust radiomics signature. However, all feature normalizations have been obtained by using a training set from each center, meaning that, if an additional center will be added, a training phase should be performed to extract normalization parameters using at least 20 patients.

This study has limitations. First, the lack of the assessment of feature reproducibility when ROIs are segmented by different operators or at different timepoints. Nevertheless, we took into account both inter- and intra-reader variability by using six ROIs with the same dimensions on the same tissue. Second, no normalization methods based on healthy tissue intensities were included in our analysis, even if they were demonstrated as useful in reducing feature differences (Isaksson *et al* 2020). However, this kind of approach requires the manual segmentation of the reference areas, which is unfeasible for large and multicenter studies and could not bridge the gap between research and clinical applications.

Even if our results provide several insights on the reduction of the feature differences among centers using some normalization methods, further analyses could be carried out, even tailoring on different clinical tasks related to the abdominal area.

5. Conclusions

In this study we evaluated the impact of normalization on radiomics features extracted from an abdominal multicenter MRI dataset involving three different centers and four different scanners. From our findings it emerged that some feature normalization methods could substantially improve the feature reproducibility, while image ones have almost no impact or, in some cases, they may worsen it. However, the image normalization methods, which may reduce the histogram distributions variability, could be a useful step when developing deep learning models. Despite we demonstrated that it exists a subgroup of reproducible features, we recommend to carefully select the proper normalization method also depending on the intended classification task to be achieved and the effect on the prognostic power of the features.

Acknowledgments

This work was funded by AIRC 5xmille Special Program Molecular Clinical Oncology - Ref. 9970, FPRC 5xmille 2013 Ministero della Salute, and FPRC 5xmille 2015 Ministero della Salute (STRATEGY), Fondazione AIRC under 5 per Mille 2018—ID. 21091 program—P.I.Bardelli Alberto, G.L. Regge Daniele.

Data availability statement

The ethical committees which approved the study did not grant permission to publish the data used in the study. The data that support the findings of this study are available upon reasonable request from the authors.

Conflict of interest statement

The authors have no relevant conflicts of interest to disclose.

ORCID iDs

Valentina Giannini https://orcid.org/0000-0001-5052-8231
Jovana Panic https://orcid.org/0000-0002-6620-5610
Daniele Regge https://orcid.org/0000-0001-8267-5279
Gabriella Balestra https://orcid.org/0000-0003-2717-648X
Samanta Rosati https://orcid.org/0000-0003-0620-594X

References

Alrahawy M, Aker M, Issa M, Ali O, Noureldin K, Gaber A, Mahgoub A, Ahmed M, Yousif M and Zeinaldine A 2022 Textural analysis as a predictive biomarker in rectal cancer *Cureus* **14** e32241

Beets-Tan R G H *et al* 2018 Magnetic resonance imaging for clinical management of rectal cancer: updated recommendations from the 2016 European society of gastrointestinal and abdominal radiology (ESGAR) consensus meeting *Eur. Radiol.* **28** 1465–75

Bobak C A, Barr P J and O'Malley A J 2018 Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales *BMC Med. Res. Methodol.* **18** 1–11

Breiding M J 2014 Computational radiomics system to decode the radiographic phenotype *Physiol. Behav.* **63** 1–18

Buch K, Kuno H, Qureshi M M, Li B and Sakai O 2018 Quantitative variations in texture analysis features dependent on MRI scanning parameters: a phantom model *J. Appl. Clin. Med. Phys.* **19** 253–64

Campello V M *et al* 2022 Minimising multi-centre radiomics variability through image normalisation: a pilot study *Sci. Rep.* **12** 1–10

Carré A *et al* 2020 Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics *Sci. Rep.* **10** 1–15

Chatterjee A, Vallieres M, Dohan A, Levesque I R, Ueno Y, Saif S, Reinhold C and Seuntjens J 2019 Creating robust predictive radiomic models for data from independent institutions using normalization *IEEE Trans. Radiat. Plasma Med. Sci.* **3** 210–5

Crombé A, Kind M, Fadli D, Le Loarer F, Italiano A, Buy X and Saut O 2020 Intensity harmonization techniques influence

radiomics features and radiomics-based predictions in sarcoma patients *Sci. Rep.* **10** 1–13
Curtin F and Schulz P 1998 Multiple correlations and bonferroni's correction *Biol. Psychiatry* **44** 775–7
Da-Ano R, Visvikis D and Hatt M 2020 Harmonization strategies for multicenter radiomics investigations *Phys. Med. Biol.* **65** 24TR02
Dikaios N *et al* 2015 Logistic regression model for diagnosis of transition zone prostate cancer on multi-parametric MRI *Eur. Radiol.* **25** 523–32
Engelhard K, Hollenbach H P, Deimling M, Kreckel M and Riedl C 2000 Combination of signal intensity measurements of lesions in the peripheral zone of prostate with MRI and serum PSA level for differentiating benign disease from prostate cancer *Eur. Radiol.* **10** 1947–53
Fortin J P *et al* 2018 Harmonization of cortical thickness measurements across scanners and sites *Neuroimage* **167** 104–20
Fusco R *et al* 2022 Radiomics in medical imaging: pitfalls and challenges in clinical management *Jpn. J. Radiol.* **40** 919–29
Giannini V, Vignati A, Mirasole S, Mazzetti S, Russo F, Stasi M and Regge D 2016 MR-T2-weighted signal intensity: a new imaging biomarker of prostate cancer aggressiveness *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **4** 130–4
Gillies R J, Kinahan P E and Hricak H 2016 Radiomics: images are more than pictures, they are data *Radiology* **278** 563–77
Granzier R W Y *et al* 2022 Test–retest data for the assessment of breast MRI radiomic feature repeatability *J. Magn. Reson. Imaging* **56** 592–604
Haralick R M, Dinstein I and Shanmugam K 1973 Textural features for image classification *IEEE Trans. Syst. Man Cybern. SMC-* **3** 610–21
Horng H, Singh A, Yousefi B, Cohen E A, Haghighi B, Katz S, Noël P B, Shinohara R T and Kontos D 2022 Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects *Sci. Rep.* **12** 1–12
Isaksson L J *et al* 2020 Effects of MRI image normalization techniques in prostate cancer radiomics *Phys. Medica* **71** 7–13
Kociołek M, Strzelecki M and Obuchowicz R 2020 Does image normalization and intensity resolution impact texture classification? *Comput. Med. Imaging Graph.* **81** 101716
Koo T K and Li M Y 2016 A guideline of selecting and reporting intraclass correlation coefficients for reliability research *J. Chiropr. Med.* **15** 155–63
Lambin P *et al* 2012 Radiomics: Extracting more information from medical images using advanced feature analysis *European Journal of Cancer* **48** 441–6
Lambin P *et al* 2017 Radiomics: the bridge between medical imaging and personalized medicine *Nat. Rev. Clin. Oncol.* **14** 749–62
Li Y, Ammari S, Balleyguier C, Lassau N and Chouzenoux E 2021 Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features *Cancers (Basel).* **13** 1–22
Lu L, Liang Y, Schwartz L H and Zhao B 2019 Reliability of radiomic features across multiple abdominal ct image acquisition settings: a pilot study using acr ct phantom *Tomography* **5** 226–31
Ly J, Minarik D, Edenbrandt L, Wollmer P and Trägårdh E 2019 The use of a proposed updated EARL harmonization of 18F-FDG PET-CT in patients with lymphoma yields significant differences in Deauville score compared with current EARL recommendations *EJNMMI Res.* **9** 0–6
Mchugh D J, Porta N, Little R A, Cheung S, Watson Y, Parker G J M, Jayson G C and Connor J P B O 2021 Image Contrast, Image Pre-Processing , and T 1 Mapping Affect Cancer Liver Metastases *Cancers (Basel)* **13** 240
Nachar N 2008 The mann-whitney U: a test for assessing whether two independent samples come from the same distribution *Tutor. Quant. Methods Psychol.* **4** 13–20
Nyú L G and Udupa J K 1999 On standardizing the MR image intensity scale *Magn. Reson. Med.* **42** 1072–81
Orlhac F, Eertink J J, Cottereau A-S, Zijlstra J M, Thieblemont C, Meignan M, Boellaard R and Buvat I 2021 A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies *Journal of Nuclear Medicine* **63** 172–9
Rai R, Holloway L C, Brink C, Field M, Christiansen R L, Sun Y, Barton M B and Liney G P 2020 Multicenter evaluation of MRI-based radiomic features: A phantom study *Med. Phys.* **47** 3054–63
Reinhold J C, Dewey B E, Carass A and Prince J L 2019 Evaluating the impact of intensity normalization on MR image synthesis *Proc. SPIE Int. Soc. Opt. Eng.* **176** 126
Scalco E, Belfatto A, Mastropietro A, Rancati T, Avuzzi B, Messina A, Valdagni R and Rizzo G 2020 T2w-MRI signal normalization affects radiomics features reproducibility *Med. Phys.* **47** 1680–91
Scalco E and Rizzo G 2017 Texture analysis of medical images for radiotherapy applications *Br. J. Radiol.* **90** 20160642
Schwier M, van Griethuysen J, Vangel M G, Pieper S, Peled S, Tempany C, Aerts H J W L, Kikinis R, Fennessy F M and Fedorov A 2019 Repeatability of multiparametric prostate MRI radiomics features *Sci. Rep.* **9** 1–16
Stamoulou E, Spanakis C, Manikis G C, Karanasiou G, Grigoriadis G, Foukakis T, Tsiknakis M, Fotiadis D I and Marias K 2022 Harmonization strategies in multicenter MRI-based radiomics *J. Imaging* **8** 303
Stanzione A, Cuocolo R, Ugga L, Verde F, Romeo V, Brunetti A and Maurea S 2022 Oncologic imaging and radiomics: a walkthrough review of methodological challenges *Cancers (Basel)* **14** 1–14
Traverso A, Wee L, Dekker A and Gillies R 2018 Repeatability and reproducibility of radiomic features: a systematic review *Int. J. Radiat. Oncol. Biol. Phys.* **102** 1143–58
Upadhaya T *et al* 2019 Comparison of radiomics models built through machine learning in a multicentric context with independent testing: identical data, similar algorithms, different methodologies *IEEE Trans. Radiat. Plasma Med. Sci.* **3** 192–200
Wang L, Mazaheri Y, Zhang J, Ishill N M, Kuroiwa K and Hricak H 2008 Assessment of biologic aggressiveness of prostate cancer: correlation of MR signal intensity with gleason grade after radical prostatectomy *Radiology* **246** 168–76
Yushkevich P A, Piven J, Hazlett H C, Smith R G, Ho S, Gee J C and Gerig G 2006 User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability *Neuroimage* **31** 1116–28
Zwanenburg A, Leger S and Vallières M L S 2016 Image biomarke standardization initiative (https://doi.org/10.1148/radiol.2020191145)