



A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study

Marie-Judith Saint Martin, Fanny Orlhac, Pia Akl, Fahad Khalid, Christophe Nioche, Irène Buvat, Caroline Malhaire, Frédérique Frouin

► To cite this version:

Marie-Judith Saint Martin, Fanny Orlhac, Pia Akl, Fahad Khalid, Christophe Nioche, et al.. A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study. Magnetic Resonance Materials in Physics, Biology and Medicine, 2020, Online ahead of print. 10.1007/s10334-020-00892-y . inserm-03029480

HAL Id: inserm-03029480

<https://inserm.hal.science/inserm-03029480>

Submitted on 28 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study

Authors: Marie-Judith Saint Martin¹, Fanny Orhac¹, Pia Akl^{1,2,3}, Fahad Khalid¹, Christophe Nioche¹, Irène Buvat¹, Caroline Malhaire^{1,3}, Frédérique Frouin¹

¹ Université Paris-Saclay, Inserm, Institut Curie, Laboratoire d'Imagerie Translationnelle en Oncologie (LITO), Bât 101B rue Henri Becquerel, 91 401 Orsay, France.

² HCL, Radiologie du groupement hospitalier Est, Hôpital Femme Mère enfant, Unité Fonctionnelle : imagerie de la femme, 3 Quai des Célestins, 69002 Lyon, France.

³ Institut Curie, Service de Radiodiagnostic, 26 rue d'Ulm, 75005 Paris, France.

Corresponding author: Marie-Judith Saint Martin, email: marie-judith-astrid.saint-martin@u-psud.fr, phone: +33169867181

Abstract

Object: Quantitative analysis in MRI is challenging due to variabilities in intensity distributions across patients, acquisitions and scanners and suffers from bias field inhomogeneity. Radiomic studies are impacted by these effects that affect radiomic feature values. This paper describes a dedicated pipeline to increase reproducibility in breast MRI radiomic studies.

Materials and Methods: T1, T2, and T1-DCE MR images of two breast phantoms were acquired using two scanners and three dual breast coils. Images were retrospectively corrected for bias field inhomogeneity and further normalised using Z-score or histogram matching. Extracted radiomic features were harmonised between coils by the ComBat method. The whole pipeline was assessed qualitatively and quantitatively using statistical comparisons on two series of radiomic feature values computed in the gel mimicking the normal breast tissue or in dense lesions.

Results: Intra and inter-acquisition variabilities were strongly reduced by the standardisation pipeline. Harmonisation by ComBat lowered the percentage of radiomic features significantly different between the three coils from 87% after bias field correction and MR normalisation to 3% in the gel, while preserving or improving performance of lesion classification in the phantoms.

Discussion: A dedicated standardisation pipeline was developed to reduce variabilities in breast MRI, which paves the way for robust multi-scanner radiomic studies but needs to be assessed on patient data.

Keywords: Breast, MRI, Reproducibility, Radiologic phantom, Image processing

Acknowledgments: We thank Sophie Lassalle, radiographer manager, for her help in acquiring the data. We are grateful to the anonymous reviewers for their helpful comments.

Introduction

Radiomics is a recent field of study involving the extraction of large amounts of quantitative imaging features from radiological images [1]. These radiomic features can then feed machine learning methods to build predictive models that might assist diagnosis and patient monitoring. Radiomic studies in breast cancer patients have shown promises, for instance in assessing the risk of breast cancer recurrence [2], detecting malignant from benign lesions [3], or estimating disease free survival [4]. Over the last few years, several attempts to predict the response to neoadjuvant chemotherapy using radiomics have been reported, but this remains a challenging task [5–10].

Radiomic studies and subsequent machine learning approaches require a substantial number of images to achieve relevant performance, which encourages the use of multicentre and retrospective data. However, many articles have highlighted the influence of scanner parameters on radiomic features in PET and CT imaging (see for instance the review [11]) and in MRI [12–15]. This so-called “scanner effect” requires standardisation and harmonisation procedures. In particular, MRI radiomic feature values have been shown to depend on magnetic field strengths, voxel size, pulse sequence parameters or receiver coils [16, 17]. The standardisation process is especially important in MR as images are expressed in arbitrary units that vary between patients, acquisitions and scanners. MR images also suffer from MR bias field non-uniformity, generating regional and local spatial inhomogeneities. Since the impact of this latter effect goes beyond the field of radiomics and affects tasks such as segmentation, an abundant literature already addresses this issue, but studies are mainly oriented towards brain MRI. Several methods have been developed to correct bias field inhomogeneity retrospectively [18]. Borys et al. [19] compared a subset of these approaches on breast phantoms. They found that the N4 algorithm [20] gave the most uniform results, slightly outperforming F3CM [21], but hinted that adapting the parameters of the method specifically for breast imaging could improve the correction. Following bias field correction, MR normalisation techniques have been applied to reduce inter-patient variabilities, the most frequent being the Z-score standardisation [8]. Shinohara et al. [22] designed a new linear approach, the hybrid White-Stripe, using white matter as a reference tissue in the brain to normalise images, from which

Fortin et al. [23] derived the voxel-based RAVEL method. Other non-linear normalisation techniques have been proposed such as histogram matching (referred as HM) by Nyul et Udupa [24], further adapted in a multiple sclerosis study by Sha et al. [25]. Bias field correction and intensity normalisation have been shown to improve the radiomic characterisation of tumours from single centre MR images of paediatric brain tumours [26] and lung cancer [27].

Recent works in glioblastoma [28, 29] and prostate cancer [16, 30] patients specifically investigated the influence of bias field correction, noise reduction and histogram normalisation on the scanner effect affecting MR radiomic features. Authors identified small subsets of features that were reproducible across scanners after standardisation but did not manage to successfully harmonise all features. A harmonisation method called ComBat initially developed to mitigate batch effects in genomic studies [31] was successfully applied to compensate for the scanner effect in PET [32], CT [33] and MR [34, 35].

Few breast radiomic studies mention bias field correction [36] or MR normalisation [8] before computing features. In this study, we propose and validate a radiomics pipeline dedicated to breast MRI. First, a bias field correction method was adapted for breast MR images to overcome the limitations of the conventional approaches. Second, two MR image standardisation techniques (Z-score, HM) were investigated to study their impact on MR intensity distribution. Third, the ComBat method was proposed to further reduce the scanner effect affecting radiomic features. Our experimental study was conducted using two breast phantoms designed for biopsy training in order to monitor the effects of standardisation and harmonisation without any biological or tissue interference. MR acquisitions following the standard clinical protocol in our institution were performed using two MR scanners and three dedicated breast coils. Our pipeline efficacy was assessed by studying the reproducibility of each radiomic feature across thirty regions in the gel mimicking normal tissue and by comparing the performance of dense lesion classification in the phantoms before and after harmonisation using ComBat.

Materials and Methods

Phantom

In order to remain as close as possible to clinical settings and to better investigate MR bias field considering the symmetry inherent to breast imaging, all the experiments described were carried out using two phantoms simultaneously. Two Multi-Modality Breast Biopsy and Sonographic” phantoms, CIRS reference 073 (Norfolk, VA, USA), were used (Fig. 1a). They consist of an elastomer membrane simulating the skin and subcutaneous fat layer of breast in patients and are filled with a branded gel (Zerdine®). Five to ten cystic lesions (5-10 mm) and ten to fifteen dense lesions (5-10 mm) are included in the gel. Half of the dense lesions are spheres including microcalcifications (Fig. 1c) while the other half are spiculated lesions (Fig. 1d). This model is dedicated to biopsy training and accurately reproduces breast tissue with lesions for MR imaging.

Image acquisition

Images were acquired in the three clinical imaging settings used in our institution. The two phantoms were scanned in a first setting, using a 1.5 T magnet, Optima MR450w (GE, MA, USA) with an 8-channel breast coil further referred to as “Coil 1”. They were also scanned in a second setting, using a 1.5T magnet, MAGNETOM Aera (Siemens, Munich, Germany) with an 18-channel breast coil (“Coil 2”). The third setting consisted in using a 16-channel Sentinelle breast coil, dedicated to diagnosis and MR-guided biopsy, on the 1.5T MAGNETOM Aera (Siemens, Munich, Germany) (“Coil 3”). For each setting, two acquisitions were acquired (Acq. A and Acq. B), between which the positions of the two phantoms were switched while the dual breast coils were unchanged.

The phantoms were scanned with three sequences routinely used in breast clinical imaging protocol to get T1-weighted, fat-saturated T2-weighted and T1-weighted DCE images, with parameters listed in Table 1. Scanning parameters slightly differed for each coil hence adding another source of variability to the settings. In clinical studies, T1 sequence is acquired for anatomical purposes and does not include fat saturation, while the functional T1-DCE sequence is acquired with high resolution voxel size, fat saturation, and is repeated five times every 90 seconds. However, only one dynamic is presented throughout this work since there was no injection of contrast agent in the phantom. Spoiled gradient recalled T1-DCE and T1 (on the GE Scanner only) images were acquired, while turbo spin echo acquisitions were performed in other cases. Fat saturation was achieved using either Dixon or SPAIR techniques. ARC (GE scanner) and GRAPPA (Siemens

scanner) multi-coil parallel imaging were used for T1-DCE acquisitions. The acquisition time including the three sequences was around 18 minutes.

In total, six 3D images (acquisitions A and B with three sequences each) were acquired for each of the three clinical settings, yielding a total of 18 raw 3D images.

	T1			Fat-saturated T2			T1-weighted DCE		
	Coil 1	Coil 2	Coil 3	Coil 1	Coil 2	Coil 3	Coil 1	Coil 2	Coil 3
TR (ms)	6.9	592	545	5544	3310	6400	6.81	5.2	5.2
TE (ms)	4.2	13	13	90	88	88	3.3	2.4	2.4
Slice thickness (mm)	1.6	3.5	3	3	3.5	3	1	0.9	0.9
Spacing between slices (mm)	0.8	4.2	3.6	3.3	4.2	3.6	1	0.9	0.9
Pixel spacing (mm)	0.68 x 0.68	0.71x 0.71	0.71 x 0.71	0.68 x 0.68	0.7 x 0.7	0.7 x 0.7	0.82 x 0.82	0.91 x 0.91	0.91 x 0.91
Pixel bandwidth (Hz/pixel)	244	130	130	558	315	375	434	355	355
Flip angle	20	148	148	160	150	180	15	10	10

Table 1 T1, T2 and T1 DCE sequence parameters for the three settings

Bias field correction

Images were corrected for bias field inhomogeneity using the SimpleITK N4BiasFieldCorrection Image filter class adapted for python from the implementation of the N4 algorithm [20] in the ITK library. The N4 method is based on the following image model:

$$I_{cor}(x) = I(x)B(x) + \eta(x),$$

where x is a voxel, I_{cor} is the corrupted image, B the bias field, I the bias-free image and η an independent Gaussian noise. In a noise-free case, using logarithmic transformation, with $\hat{I} = \log I$:

$$\widehat{I_{cor}} = \hat{I} + \hat{B}$$

The N4 method uses an iterative multi-scale optimisation approach, at iteration n

$$\hat{I}^n = \hat{I}^{n-1} - \widehat{B_{res}}^n$$

$$\hat{I}^n = \hat{I}^{n-1} - S^*\{\hat{I}^{n-1} - E\{\hat{I}|\hat{I}^{n-1}\},$$

where S^* is an adapted B-spline approximation, $\hat{I}^0 = \widehat{I_{cor}}$ and B_{res} is the residual bias field at step n .

The number of resolution levels and the number of iterations at each level are set by default to four levels and fifty iterations but can be changed. A default mask to select the pixels used to estimate the bias field is defined using Otsu thresholding unless a specific mask is provided.

The impact of the hyper parameters mentioned above were investigated by running several trials using three, four, five or six resolution levels, fifty or a hundred iterations, combined with either the default mask or a full mask of the phantom.

To assess the ability of the N4 correction to reduce intensity non-uniformities within similar tissue types, voxels corresponding to the background gel and embedded masses were clustered using the k-means algorithm. The clustering results were compared before and after the bias field correction. As the inner part is made of three different materials (background gel, dense masses and cyst masses), that have different physical properties, the number of clusters was set to three.

The performance of N4 algorithm was also assessed by comparing the coefficients of variation of the mean intensity of small regions drawn in the background gel for the different corrections: fifteen 3D spherical regions of 600 voxels each were drawn using the LIFEx freeware [37] (www.lifexsoft.org) on every raw acquisition. These spheres were located in the background neutral gel of the phantom, avoiding any cyst or dense masses. As there were two acquisitions per coil, regions from the same coil were pooled to get thirty regions per coil.

MR normalisation

MR images were normalized after bias field correction as it has been shown that pre-correcting intensity non-uniformities leads to an improved standardisation [38].

Two types of normalisation were performed separately and compared: 1) Z-score standardisation using a mask of the phantom to compute the mean and standard deviation of intensities (linear transform); 2) piecewise linear histogram matching [24, 25]. Histogram matching includes two stages: first, HM learns landmarks of a standard histogram and then landmarks of the image histograms are non-linearly mapped to the ones of the standard histogram to align the intensity distributions. HM was applied independently on the three sequences with codes adapted from Reinhold et al. [39], using the decile landmarks and standard scale defined by Shah et al. [25].

The impact of normalisation in correcting inter-subject and inter-coil variabilities was evaluated by qualitatively comparing intensity histogram alignment and by using texture analyses.

Radiomic analysis

After MR normalisation, four MR volumes of the same acquisition were available, corresponding to raw data, N4-corrected data, z-score normalised-N4-corrected data, and HM normalised-N4-corrected data.

All 18x4 (3 sequences x 3 coils x 2 acquisitions x 4 transformations) MR volumes were resampled, as recommended by Image Biomarker Standardisation Initiative guidelines [40] before extracting features, using nearest neighbour interpolation: T1 and fat-saturated T2 images were resampled to 0.7x0.7x4 mm³ voxels and T1-weighted DCE images to 1x1x1 mm³ voxels.

In addition to the regions inside the background gel, dense lesions were segmented on N4-corrected data using the k-means algorithm, these latter regions being manually corrected, when it was necessary.

For each MR volume and each region, 42 radiomic features were computed with LIFEx v5.79 [37] in compliance with the Image Biomarker Standardisation Initiative guidelines [40]. Besides first order features, texture features included indices from the grey-level co-occurrence matrix (GLCM), the grey-level run length matrix (GLRLM), the grey-level zone length matrix (GLZLM) and the neighbourhood grey-level different matrix (NGLDM). The list of radiomic features is provided in Supplemental Table 1. For texture calculations, absolute discretization was chosen [26]. For each sequence and each step of the standardisation pipeline separately, the minimum and maximum intensities inside the regions were calculated to determine the range of intensities and the average standard deviation of intensities over the regions was defined as the fixed bin size.

Harmonisation of radiomic features

Since radiomic feature values might differ between the three experimental settings even after the different processing steps, for every sequence separately, the distributions of the radiomic features extracted after normalisation were harmonised across the three coils using the ComBat method [32-34]. The ComBat method aims at correcting any differences that could be due to coils, scanners and/or scanning parameters [41, 42]. For each radiomic feature y measured in region j in centre i , the scanner effect on feature y_{ij} can be modelled as:

$$y_{ij} \sim \alpha + \gamma_i + \delta_i \varepsilon_{ij}$$

where α is the overall value of the radiomic feature y , γ_i is an additive centre effect and $\delta_i \varepsilon_{ij}$ a multiplicative centre effect associated to an error term.

The ComBat method corrects the distributions by calculating $\hat{\alpha}$, $\hat{\gamma}_i$ et $\hat{\delta}_i$ as estimators of α , γ_i et δ_i using maximum likelihood estimation so that:

$$y_{ijcorrected} = \frac{y_{ij} - \hat{\alpha} - \hat{\gamma}_i}{\hat{\delta}_i} + \hat{\alpha} ,$$

The non-parametric form of the method was used without any empirical Bayes assumption. Using R codes by Fortin et al. [41, 42], ComBat was applied on the feature values measured in the gel and in the dense lesions separately. A specific transformation was determined for every feature independently.

Statistical analysis

Statistical analyses were performed in R, and p -values less than 0.05 were interpreted as statistically significant.

As radiomic features are computed in 30 regions inside the background gel, they should be comparable. For each step of the pipeline and after harmonisation by ComBat, inter-coil differences in the statistical distributions of each radiomic feature were assessed using the Kruskal-Wallis test. A total of 3 (T1, T2 and T1-DCE sequences) x 5 (raw, N4-corrected, z-score normalised-N4-corrected, HM normalised-N4-corrected, HM normalised-N4-corrected after ComBat) x 42 (radiomic features) Kruskal-Wallis tests were thus performed.

To provide a synthetic view of the test results, five ranges of p -values were defined: $p < 10^{-5}$, $10^{-5} \leq p < 10^{-3}$, $10^{-3} \leq p < 0.01$, $0.01 \leq p < 0.05$ and $p \geq 0.05$. Radiomic features were then classified into the five ranges defined above depending on the p -value of their Kruskal-Wallis test. For each sequence at each step of the pipeline, the number of features in every range was reported.

To check that the ComBat method does not reduce the discriminative power of radiomic features, we evaluated the impact of ComBat on the task of separating the two types of dense lesions in the T1, T2 and T1-DCE sequences. The 42 radiomic features computed from the lesions in the HM-normalised-N4-corrected images were compared before and after harmonisation by ComBat using Wilcoxon tests.

Results

Bias field correction

Default parameters of the N4 algorithm (four resolution levels, fifty iterations per level and the use of a mask defined by Otsu thresholding) proved suboptimal for breast MR images. From a qualitative point of view, the corrected images showed little improvement when compared to the raw images (Fig. 1b and 1g). The bias field estimated with the default parameters was almost flat on the upper half of the images (Fig. 1f). These upper regions are the regions of interest where clinical information will be looked for, whereas lower regions have less clinical relevance (corresponding to zones posterior to the thorax in patients). Increasing the number of iterations per level did not improve the corrections as the results seemed to stabilise after fifty iterations. Running the algorithm with five resolution levels instead of four yielded a bias field with much stronger variations in the upper zones resulting in a greater correction of the images. Combining it with a full mask of the phantoms to estimate the bias field further improved the correction in the upper regions of interest (Fig. 1m).

Fig. 1 **a** Phantom. **b** Raw T1 image from Coil 3. **c** Dense lesion with microcalcification. **d** Dense spiculated lesion. **e** Default mask. **f** Bias field estimated with mask e and 4 fitting levels. **g** Corrected image obtained from bias field f. **h** Default mask. **i** Bias field estimated with mask h and 5 fitting levels. **j** Corrected image from bias field

i. **k** Full mask. **l** Bias field estimated with mask **k** and 5 fitting levels. **m** Corrected image from bias field **l**. Red arrows point at regions with residual intensity non-uniformity

The impact of the N4 correction on the coefficients of variation of the means over the regions across coils in different correction scenarios is shown in Fig. 2.

Fig. 2 a Coefficients of variation of the means over thirty regions across settings with different corrections. **b** Example of 3 regions (in blue, light blue and yellow) drawn in LIFEx

The effect of N4 correction with the full mask, 5 levels (50 Iterations) on k-means segmentation results using three clusters on a raw image versus the image after bias field correction is shown in Fig. 3.

Fig. 3 a, d T1-weighted DCE image from coil 3. **b, e** Histogram of the inner layer voxels of image coloured by the results of k-means clustering. **c, f** k-means clustering results overlaid on image. First line: raw image. Second line: N4 corrected (full mask, 5 levels, 50 iterations) image

Fig. 4 presents all bias field estimations across sequences and acquisitions, with images normalised so that the mean intensity in the mask used to estimate the bias field is equal to 1.

Fig. 4 Examples of estimated bias fields across sequences and acquisitions

MR normalisation

Histograms of image intensities within the phantoms (as defined by the mask used for N4 correction) are shown in Fig. 5 for the four stages of the post-processing pipeline.

The different post-processing methods had a similar behaviour across the three sequences. In raw images, the peaks of the histograms were not aligned before any correction. Intensities from coil 1 (the GE machine), in particular, spread on a significantly greater range than the intensities from the two Siemens coils. N4 correction

sharpened the peaks but did not align them. Z-score normalisation combined with N4 correction realigned perfectly acquisitions from the same coils and managed to align the peaks of different coils around the same value. The alignment was nevertheless not optimal, especially in high intensities in T2 images. Histogram matching produced the best alignments whatever the coils.

Fig. 5 Image intensity histograms of the six acquisitions (Acq.) for the four steps of the standardisation pipeline across the three sequences. Each row represents a sequence and each column a step of the pipeline

Harmonisation of radiomic features

To illustrate the impact of the pipeline on radiomic feature values, Fig. 6 shows the statistical distributions of the Short-Zone High Gray-level Emphasis (GLZLM_SZHGE) feature extracted from regions on fat-saturated T2 images across coils for the four stages of the standardisation pipeline, and after further harmonisation using ComBat. This example shows that the Z-score and HM normalisation contributed to realigning the distributions across coils (Fig. 6c, 6d) but that further harmonisation using ComBat was needed to co-align all three coils distributions (Fig. 6e). Fig. 6f presents a plot of the ComBat corrected values against the values before harmonisation to illustrate the three transformations (one per coil) applied to the feature.

Fig. 6 Statistical distributions across coils of the GLZLM_SZHGE texture feature extracted from **a** raw T2 images. **b** N4 corrected (full mask, 5 levels) T2 images. **c** Z-score normalised-N4 corrected T2 images. **d** Histogram-matched-N4 corrected T2 images. **e** Histogram-matched-N4 corrected T2 images and harmonised by ComBat. **f** Affine transformations (one per coil) of GLZLM_SZHGE values estimated by ComBat

Table 2 reports the number of features for which Kruskal-Wallis p -values are inside a specific range. For instance, in T1 raw images, 37 out 42 features were significantly different between the 3 coils with a p -value $p < 10^{-5}$, 3 features were significantly different with a p -value $10^{-5} \leq p < 10^{-3}$, 1 with a p -value

$10^{-3} \leq p < 0.01$, another one with a p -value $0.01 \leq p < 0.05$ and no feature yielded a p -value greater than 0.05. Detailed results are given for each radiomic feature in Supplemental Tables 2, 3 and 4.

The same trend was observed across T1 and T2 sequences: the number of features that were significantly different ($p < 0.05$) decreased gradually when they were computed from N4-corrected data, Z-score-normalised N4-corrected data, HM-normalised N4-corrected data, ComBat harmonised HM-normalised N4-corrected data and the number of smaller p -values was reduced accordingly. Harmonisation by ComBat was essential to reduce drastically the number of significantly different features in all three settings, especially for T1-weighted DCE features.

T1	P value	Raw images	N4 correction	N4 & Z-score	N4 & HM	N4 & HM & ComBat
	$p < 10^{-5}$	37	37	24	5	0
	$10^{-5} \leq p < 10^{-3}$	3	1	6	7	0
	$10^{-3} \leq p < 0.01$	1	3	3	11	0
	$0.01 \leq p < 0.05$	1	0	3	9	2
	$0.05 \leq p$	0	1	6	10	40
T2	P value	Raw images	N4 correction	N4 & Z-score	N4 & HM	N4 & HM & ComBat
	$p < 10^{-5}$	38	36	26	13	0
	$10^{-5} \leq p < 10^{-3}$	4	2	5	6	0
	$10^{-3} \leq p < 0.01$	0	1	5	10	0
	$0.01 \leq p < 0.05$	0	2	2	4	1
	$0.05 \leq p$	0	1	4	9	41
T1 weighted DCE	P value	Raw images	N4 correction	N4 & Z-score	N4 & HM	N4 & HM & ComBat
	$p < 10^{-5}$	39	39	39	40	0
	$10^{-5} \leq p < 10^{-3}$	1	0	1	1	0
	$10^{-3} \leq p < 0.01$	1	1	0	0	0
	$0.01 \leq p < 0.05$	0	0	0	0	1
	$0.05 \leq p$	1	2	2	1	41

Table 2 Number of radiomic features in the 5 different ranges of p -values. The p -values correspond to Kruskal-Wallis tests between the three coils, radiomic features being extracted from 30 similar regions. Results are given for the three MR sequences and each main step of the processing pipeline

Regarding the two types of dense lesions, Wilcoxon tests were performed for each feature and each sequence. Before harmonisation, on T1 (respectively T2, T1-DCE) images, 10 (respectively 39, 7) features out of 42 were significantly different between the two lesion types, whereas after ComBat harmonisation, 32 (respectively 39, 21) features were significantly different. Fig. 7 shows the impact of ComBat on the mean intensity, with reduced inter-scanner effect.

Fig. 7 Mean intensity before and after ComBat harmonisation across sequences and coils of lesion 1 (Dense lesion with microcalcification) and lesion 2 (Dense spiculated lesion). Asterisks denote cases where the difference between the two lesion types is significant.

Discussion

The present study suggests that standardisation methods developed for brain or lung MRI should be adapted specifically to breast MR images. The process we propose includes a bias field correction to reduce local/regional inhomogeneities in similar regions (intra-image variabilities), an intensity normalisation to decrease inter-acquisition variabilities, and statistical harmonisation to make results across coils comparable. We have shown that the three steps, each tackling a different source of variability, are all needed and complementary. They pave the way towards an efficient standardisation pipeline for multi-scanner radiomic studies of patients' acquisitions.

To enable retrospective patients' studies, bias field correction was based on an *a posteriori* method. Comparisons of bias fields with different settings of the N4 algorithm led to a set of parameters appropriate for breast MRI when using dual breast coils. Based on our study, we recommend using a mask including the internal part of the breast phantoms (unlike the mask defined by Otsu's threshold) and performing the optimisation across five resolution levels (instead of four) with fifty iterations per level. Using the default parameters optimised for brain MRI underestimated the variations in the bias field, even when using the mask including the phantom inner part. It resulted in intensity non-uniformity inside this mask, where MR information is of prime importance in a clinical context (Fig. 1g). The drawback of the default mask and five resolution levels was also illustrated in Fig. 1. The bias field was indeed underestimated in the central part of the phantom, yielding a hypersignal effect in the corrected image (Fig. 1j) and thus increased heterogeneity in the background gel hence an increase in the coefficient of variation of the mean intensity (Fig. 2). The analysis of the quantitative assessments of all experiments (Fig. 2) showed that the proposed breast specific N4 parameters led to the greatest decrease, across coils, of the coefficients of variation of the mean intensity over homogeneous regions inside the phantom. The k-means clustering performed on the inner part of the phantom clearly shows how N4 correction reduced intensity variations across tissue types. In addition, the

intensity histogram of the inner part of the phantom on N4 corrected images showed a strong sharpening of its peak around the mean value of the largest structure, i.e. the background gel (Fig. 3e). The overlay of the segmented regions demonstrated a clear improvement in the identification of masses on N4 corrected images (Fig. 3f). Bias field correction thus appeared essential to improve homogeneity inside the breast MR images and is crucial for a correct segmentation of abnormalities in the breast. It should be underlined that the estimated bias fields depend not only on the MR scanner, but also on coils, type of sequence (T1, T2, T1 DCE), and on the positioning of the phantoms inside the breast coils. As shown in Fig. 4, the coil has a high impact. Indeed, bias field images from coils 2 and 3 originating from the same MR scanner were quite different. Using a same coil, bias field also showed large fluctuations across sequences. In clinical acquisitions, the N4 correction should be applied separately on each dynamic of the T1-DCE acquisition, due to the change in contrast agent concentration between the different time frames.

To reduce inter-subject and multi-scanner variabilities, MR normalisation was performed after bias field correction. Linear approaches using a reference tissue, similar to Shinohara et al. method [22] involving white matter in the brain were not reported as no satisfying reference tissue could be found in breast for all sequences, despite attempts with the subcutaneous fat layer of the breast. Studying the co-alignment of intensity histograms across acquisitions and coils highlighted the impact of intensity normalisation, and the good performance of the histogram matching approach. Results from Fig. 5 supported the idea that it was necessary to go beyond linear normalisation and Z-score standardisation [8], confirming findings by Nyul et al. [24] and Fortin et al. [23]. Z-score normalisation indeed squashed all intensities inside a range of values but did not succeed at aligning tissue-specific peaks. As observed by Isaksson et al. [43] (though with different types of landmarks) in the normalisation of prostate radiomics, the piecewise linear histogram matching gave excellent results in realigning intensity distributions. However, histogram matching depends on the set of images selected to extract a standard histogram. In a clinical setting, it will be important to use patient images coming from a wide range of scanner and biological variabilities to identify robust landmarks [25] that will perform as well on images of potential new patients included in the study.

Considering radiomic features computed inside thirty similar regions, statistical tests showed that N4 correction combined with histogram matching normalisation could not completely remove the scanner effects.

Each stage of the pipeline decreased the number of features that were significantly different between the three coils, but it was not sufficient to harmonise all radiomic features. This result agrees with the trends reported in glioblastoma [28, 29] and prostate [16, 30] cancer patients. Further harmonisation of the radiomic features is needed and ComBat succeeded in realigning feature distributions across scanners. Some studies normalise the features using scaling or z-score [44, 45] separately for each centre but unlike ComBat, these methods cannot model possible co-variables that could affect the features [31, 41, 42]. Performance in separating two different models of dense lesions was preserved (for T2 sequences) or largely improved (for T1-based sequences) by ComBat harmonisation, suggesting that ComBat successfully harmonises features across coils without removing differences mimicking biological variations. Though ComBat has a major role in reducing the scanner effect, the N4 correction and the normalisation are also needed to reduce intra-image and inter-acquisition variabilities that cannot be accurately compensated by the ComBat affine transformation. Combining the corrections is thus essential to correct for the different sources of variabilities.

The present study has several limitations. First, the CIRS model was built to be usable in multiple imaging modalities and not specifically in MRI. The phantom was also aimed at biopsy training providing lesions that could be biopsied multiple times and was therefore not designed for radiomic studies unlike phantoms used in normalisation [12, 15, 46]. The phantom was made from simple materials to capture the global breast heterogeneous appearance but not to mimic the very fine heterogeneity that could be observed in tumours and modelled in other phantoms [47]. Another limitation is that our experiments were performed using two MR scanners and three coils at the same institution, but we are confident in the possibility to extend our results to other scanners, centres and acquisitions protocols. Finally, there is always an inherent limit in using a phantom to assess performances of methods that we want to apply in clinical settings. Nevertheless, phantoms offered the opportunity to properly monitor the effects of standardisation without any interference of biological covariates.

Conclusion

This study shows the necessity to use a standardisation pipeline before performing radiomic studies involving MR breast images acquired using multiple settings. A retrospective bias field correction dedicated to dual

breast coils and non-linear MR intensity normalisation reduced the scanner effect for subsets of radiomic features, but further statistical harmonisation was needed to fully correct for it. The results were obtained on breast phantoms and future work will assess the pipeline on patient data, where biological and pathological variations increase the sources of MR intensity variations.

Authors' Contributions:

Saint Martin: Conception and study design, Analysis and interpretation of data, Drafting of manuscript, Critical revision

Orlhac: Conception and study design, Analysis and interpretation of data, Critical revision

Akl: Acquisition of data, Analysis and interpretation of data, Critical revision

Khalid: Acquisition of data, Analysis and interpretation of data, Critical revision

Nioche: Analysis and interpretation of data, Critical revision

Buvat: Analysis and interpretation of data, Critical revision

Malhaire: Conception and study design, Acquisition of data, Critical revision

Frouin: Conception and study design, Analysis and interpretation of data, Critical revision

Compliance with Ethical Standards

Funding: Pia Akl was funded by 'Bourse Curie M2 2018' by Institut Curie.

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: Images are more than pictures, they are data. *Radiology* 278:563–577.
2. Li H, Zhu Y, Burnside ES, Drukker K, Hoadley KA, Fan C, Conzen SD, Whitman GJ, Sutton EJ, Net JM, Ganott M, Huang E, Morris EA, Perou CM, Ji Y, Giger ML (2016) MR imaging radiomics signatures for predicting

the risk of breast cancer recurrence as given by research versions of MammaPrint, oncotype DX, and PAM50 gene assays. *Radiology* 281:382–391.

3. Bickelhaupt S, Paech D, Kickingereder P, Steudle F, Lederer W, Daniel H, Götz M, Gähler N, Tichy D, Wiesenfarth M, Laun FB, Maier-Hein KH, Schlemmer H-P, Bonekamp D (2017) Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography. *J Magn Reson Imaging* 46:604–616.
4. Park H, Lim Y, Ko ES, Cho HH, Lee JE, Han BK, Ko EY, Choi JS, Park KW (2018) Radiomics signature on magnetic resonance imaging: Association with disease-free survival in patients with invasive breast cancer. *Clin Cancer Res* 24:4705–4714.
5. Fan M, Wu G, Cheng H, Zhang J, Shao G, Li L (2017) Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients. *Eur J Radiol* 94:140–147.
6. Braman NM, Etesami M, Prasanna P, Dubchuk C, Gilmore H, Tiwari P, Pletcha D, Madabhushi A (2017) Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res* 19:57.
7. Thibault G, Tudorica A, Afzal A, Chui SYC, Naik A, Troxell ML, Kemmer KA, Oh KY, Roy N, Jafarian N, Holtorf ML, Huang W, Song X (2017) DCE-MRI Texture Features for Early Prediction of Breast Cancer Therapy Response. *Tomogr* 3:23–32.
8. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, Sun K, Tang Z, Jiang H, Li H, Xiong Q, Ding Y, Zhao X, Wang K, Liu Z, Tian J (2019) Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study. *Clin Cancer Res* 25:3538–3547.
9. Granzier RWY, van Nijnatten TJA, Woodruff HC, Smidt ML, Lobbes MBI (2019) Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review. *Eur J Radiol* 121:108736.
10. Eun NL, Kang D, Son EJ, Park JS, Youk JH, Kim J-A, Gweon HM (2020) Texture Analysis with 3.0-T MRI for Association of Response to Neoadjuvant Chemotherapy in Breast Cancer. *Radiology* 294:31–41.
11. Traverso A, Wee L, Dekker A, Gillies R (2018) Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int J Radiat Oncol Biol Phys* 102:1143–1158.
12. Waugh SA, Lerski RA, Bidaut L, Thompson AM (2011) The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. *Med Phys* 38:5058–5066.
13. Saha A, Yu X, Sahoo D, Mazurowski MA (2017) Effects of MRI scanner parameters on breast cancer radiomics. *Expert Syst Appl* 87:384–391.
14. Buch K, Kuno H, Qureshi MM, Li B, Sakai O (2018) Quantitative variations in texture analysis features dependent on MRI scanning parameters: A phantom model. *J Appl Clin Med Phys* 19:253–264.
15. Rai R, Holloway LC, Brink C, Field M, Christiansen RL, Sun Y, Barton MB, Liney GP (2020) Multicentre Evaluation of MRI-based Radiomics Features: A Phantom Study. *Med Phys*. doi: 10.1002/mp.14173
16. Chirra P (2018) Empirical evaluation of cross-site reproducibility in radiomic features for characterizing tumor appearance on prostate MRI. In: *Progress in Biomedical Optics and Imaging - Proceedings of SPIE, SPIE-Intl Soc Optical Eng*, p 10
17. Ford J, Dogan N, Young L, Yang F (2018) Quantitative Radiomics: Impact of Pulse Sequence Parameter Selection on MRI-Based Textural Features of the Brain. *Contrast Media Mol Imaging* 2018:1729071.
18. Song S, Zheng Y, He Y (2017) A review of Methods for Bias Correction in Medical Images. *Biomed Eng Rev* 1:1.
19. Frackiewicz M, Borys D, Gorczewski K, Serafin W, Palus H, Kijonka M (2018) The evaluation of correction algorithms of intensity nonuniformity in breast MRI images: a phantom study. In: *Proceedings of the Tenth Int. Conf. Mach. Vis. (ICMV 2017), SPIE-Intl Soc Optical Eng, Vienna*, p 15
20. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging* 29:1310–1320.

21. Lin M, Chan S, Chen J-H, Chang D, Nie K, Chen S-T, Lin C-J, Shih T-C, Nalcioğlu O, Su M-Y (2010) A new bias field correction method combining N3 and FCM for improved segmentation of breast density on MRI. *Med Phys* 38:5–14.
22. Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, Jarso S, Pham DL, Reich DS, Crainiceanu CM (2014) Statistical normalization techniques for magnetic resonance imaging. *NeuroImage Clin* 6:9–19.
23. Fortin JP, Sweeney EM, Muschelli J, Crainiceanu CM, Shinohara RT (2016) Removing inter-subject technical variability in magnetic resonance imaging studies. *Neuroimage* 132:198–212.
24. Nyúl LG, Udupa JK (1999) On standardizing the MR image intensity scale. *Magn Reson Med* 42:1072–1081.
25. Shah M, Xiao Y, Subbanna N, Francis S, Arnold DL, Collins DL, Arbel T (2011) Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Med Image Anal* 15:267–282.
26. Goya-Outi J, Orlhac F, Calmon R, Alentorn A, Nioche C, Philippe C, Puget S, Boddaert N, Buvat I, Grill J, Frouin V, Frouin F (2018) Computation of reliable textural indices from multimodal brain MRI: Suggestions based on a study of patients with diffuse intrinsic pontine glioma. *Phys Med Biol* 63:105003.
27. Lacroix M, Frouin F, Dirand A-S, Nioche C, Orlhac F, Bernaudin J-F, Brillet P-Y, Buvat I (2020) Correction for Magnetic Field Inhomogeneities and Normalization of Voxel Values Are Needed to Better Reveal the Potential of MR Radiomic Features in Lung Cancer. *Front Oncol* 10:43.
28. Um H, Tixier F, Bermudez D, Deasy JO, Young RJ, Veeraraghavan H (2019) Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. *Phys Med Biol* 64:165011.
29. Moradmand H, Aghamiri SMR, Ghaderi R (2020) Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *J Appl Clin Med Phys* 21:179–190.
30. Shiradkar R, Ghose S, Jambor I, Taimen P, Ettala O, Purysko AS, Madabhushi A (2018) Radiomic features from pretreatment biparametric MRI predict prostate cancer biochemical recurrence: Preliminary findings. *J Magn Reson Imaging* 48:1626–1636.
31. Johnson W, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118–127.
32. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I (2018) A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med* 59:1321–1328.
33. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I (2019) Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology* 291:53–59.
34. Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, Ayache N, Frouin F, Duron L, Buvat I (2020) How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol* 1–9.
35. Whitney HM, Li H, Ji Y, Liu P, Giger ML (2020) Harmonization of radiomic features of breast lesions across international DCE-MRI datasets. *J Med Imaging* 7:1.
36. Wu J, Sun X, Wang J, Cui Y, Kato F, Shirato H, Ikeda DM, Li R (2017) Identifying relations between imaging phenotypes and molecular subtypes of breast cancer: Model discovery and external validation. *J Magn Reson Imaging* 46:1017–1027.
37. Nioche C, Orlhac F, Boughdad S, Reuze S, Goya-Outi J, Robert C, Pellot-Barakat C, Soussan M, Frouin F, Frouin V, Buvat I (2018) LIFEx: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res* 78:4786–4789.
38. Madabhushi A, Udupa JK (2005) Interplay between intensity standardization and inhomogeneity correction in MR image processing. *IEEE Trans Med Imaging* 24:561–576.
39. Reinhold JC, Dewey BE, Carass A, Prince JL (2019) Evaluating the impact of intensity normalization on MR

image synthesis. In: Proceedings of Medical Imaging 2019: Image Processing, International Society for Optics and Photonics, San Diego, p 126

40. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, Bogowicz M, Boldrini L, Buvat I, Cook GJR, Davatzikos C, Depeursinge A, Desseroit MC, Dinapoli N, Dinh CV, Echegaray S (2020) The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 295:328–338.
41. Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, Adams P, Cooper C, Fava M, McGrath PJ, McInnis M, Phillips ML, Trivedi MH, Weissman MM, Shinohara RT (2018) Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167:104–120.
42. Fortin harP, Parker D, Tunç B, Watanabe T, Elliott M, Ruparel K, Roalf D, Satterthwaite T, Gur R, Gur R, Schultz R, Verma R, Shinohara R (2017) Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* 161:149–170.
43. Isaksson LJ, Raimondi S, Botta F, Pepa M, Gugliandolo SG, De Angelis SP, Marvaso G, Petralia G, De Cobelli O, Gandini S, Cremonesi M, Cattani F, Summers P, Jereczek-Fossa BA (2020) Effects of MRI image normalization techniques in prostate cancer radiomics. *Phys Med* 71:7–13.
44. Chatterjee A, Vallieres M, Dohan A, Levesque IR, Ueno Y, Saif S, Reinhold C, Seuntjens J (2019) Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization. *IEEE Trans Radiat Plasma Med Sci* 3:210–215.
45. Castaldo R, Pane K, Nicolai E, Salvatore M, Franzese M (2020) The impact of normalization approaches to automatically detect radiogenomic phenotypes characterizing breast cancer receptors status. *Cancers* 12:518.
46. Bianchini L, Botta F, Origgi D, Rizzo S, Mariani M, Summers P, García-Polo P, Cremonesi M, Lascialfari A (2020) PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis. *Phys Med* 71:71–81.
47. Valladares A, Beyer T, Rausch I (2020) Physical imaging phantoms for simulation of tumor heterogeneity in PET, CT, and MRI: An overview of existing designs. *Med Phys* 47:2023–2037.

Figure1

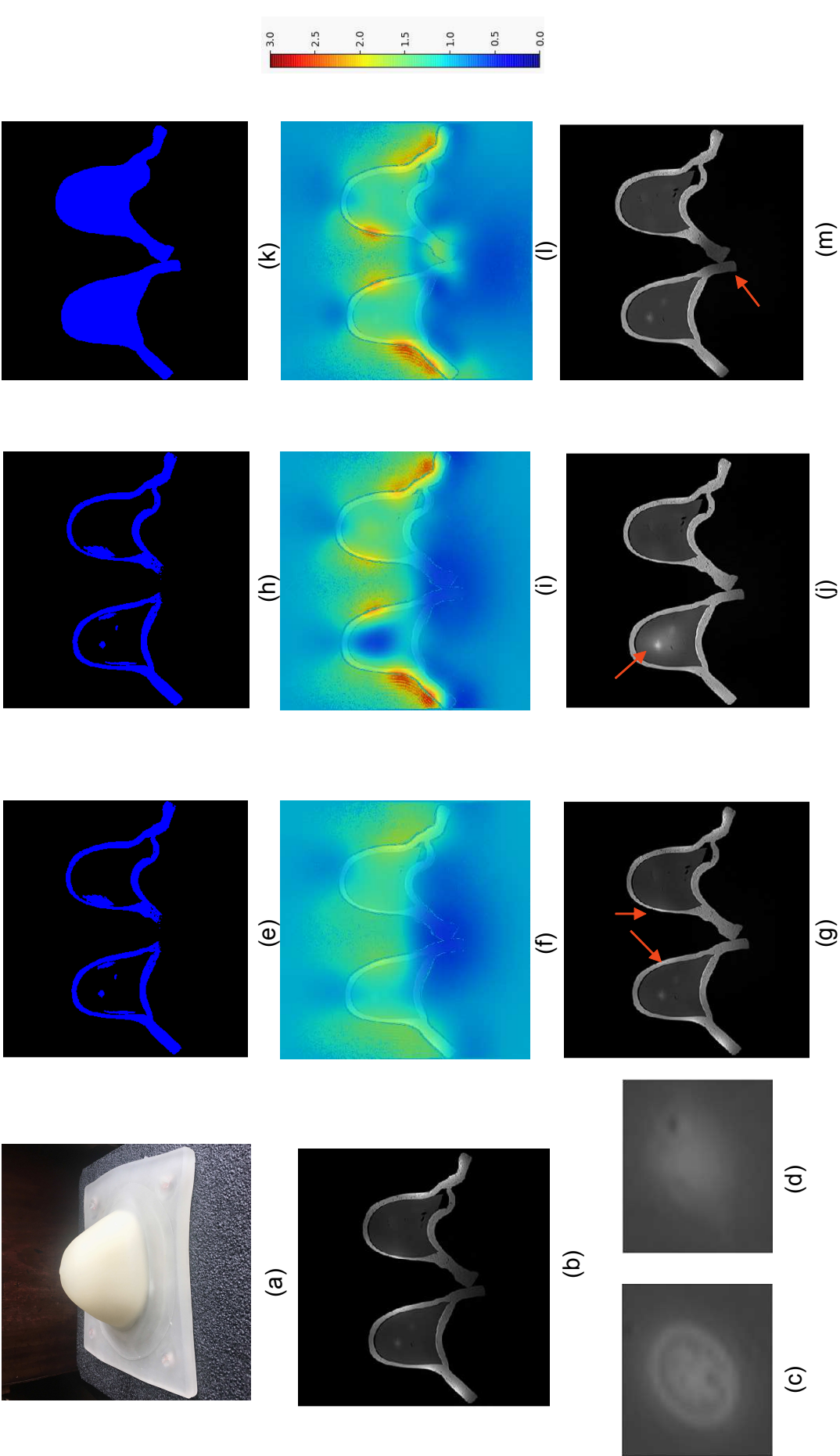


Figure2

[Click here to access/download;Figure;Fig2.eps](#)

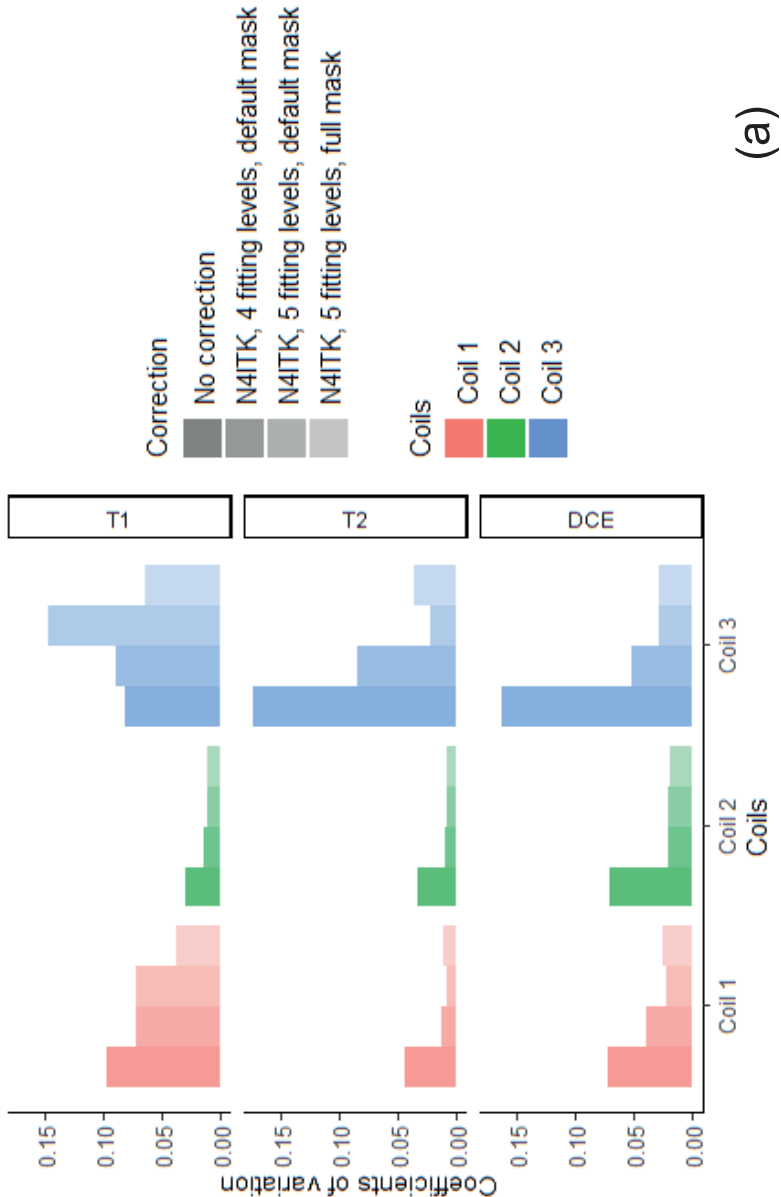
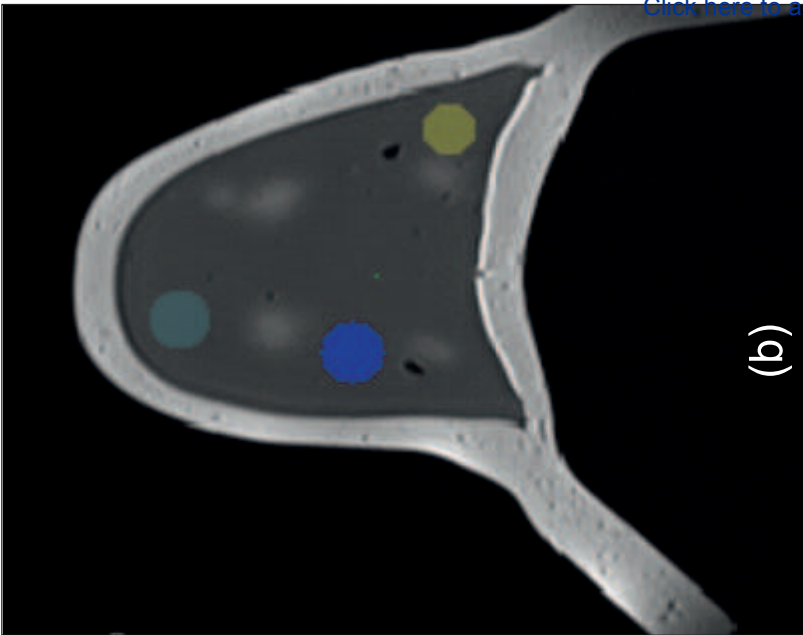


Figure3

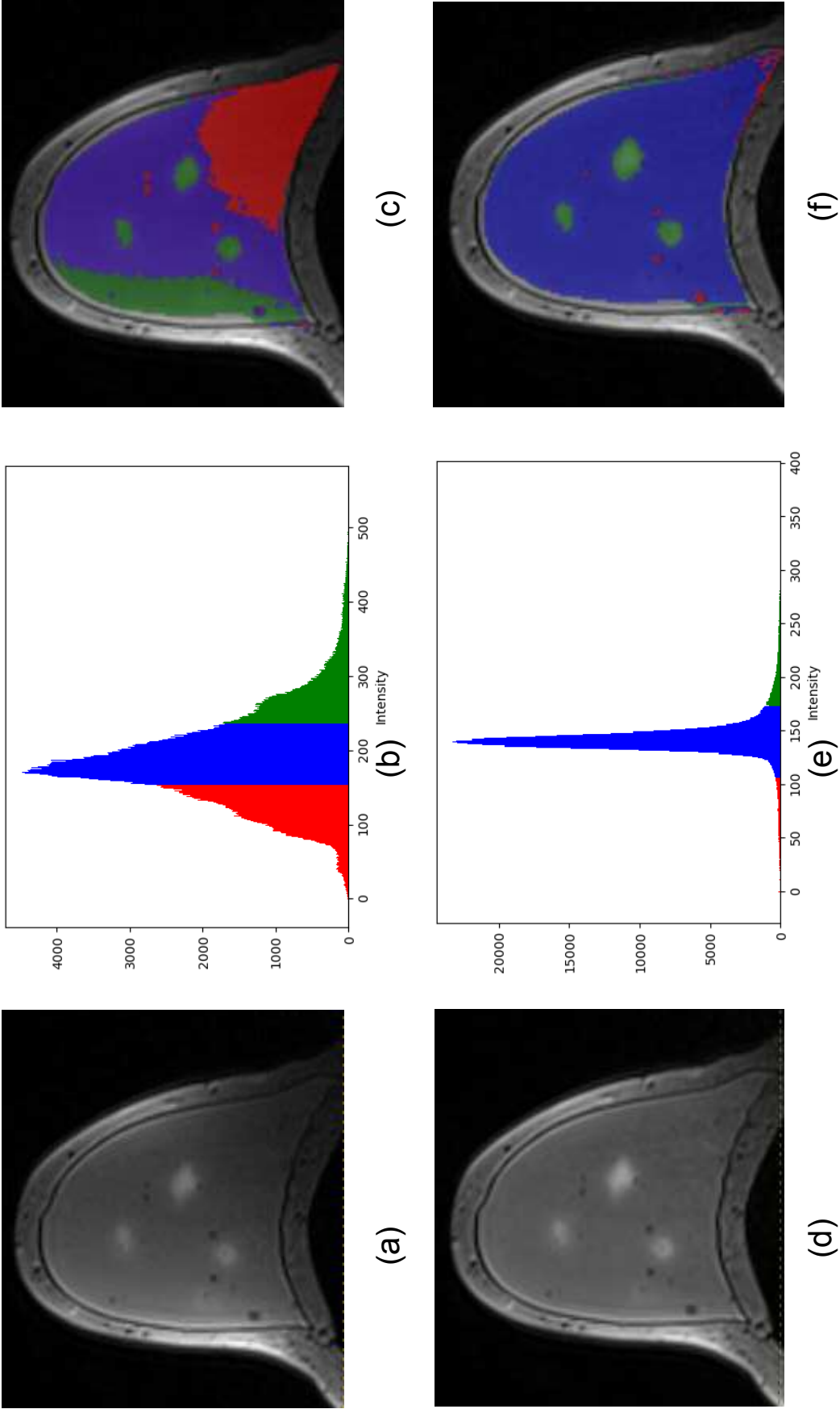


Figure4

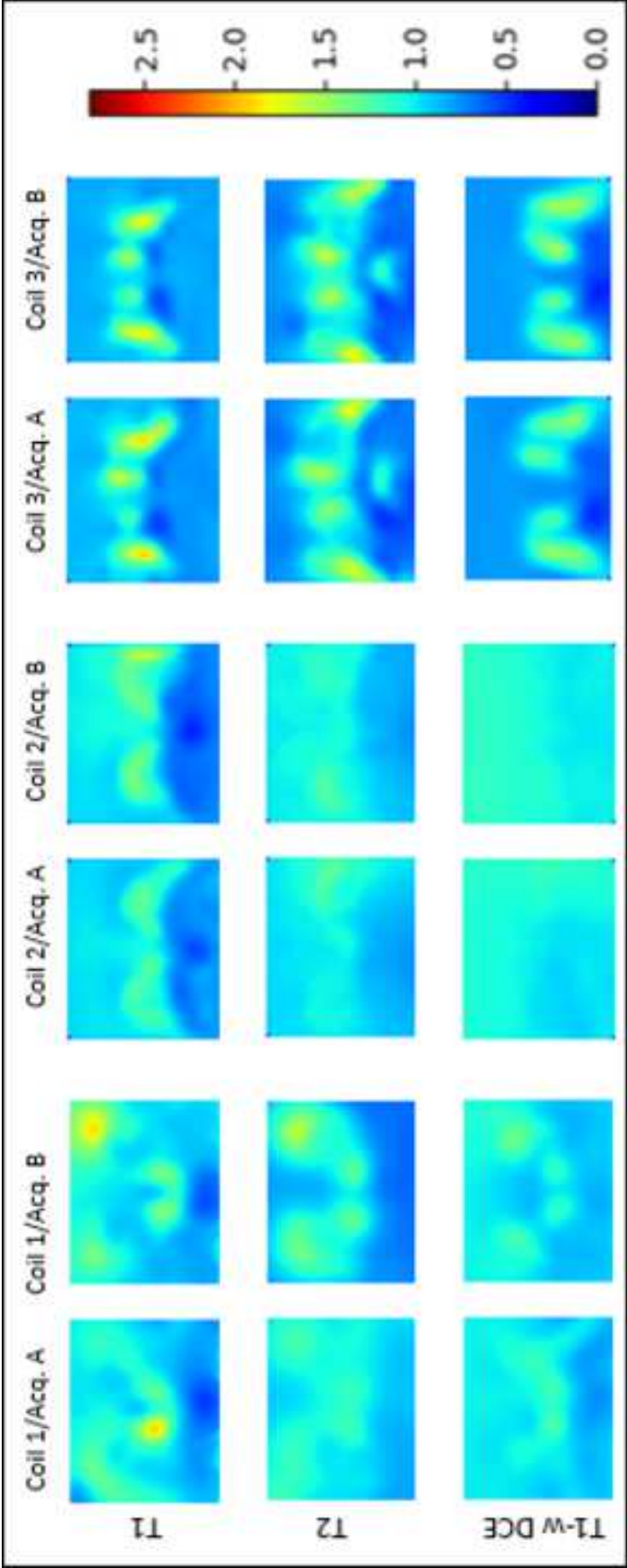


Figure5

[Click here to access/download;Figure;Fig5.eps](#)

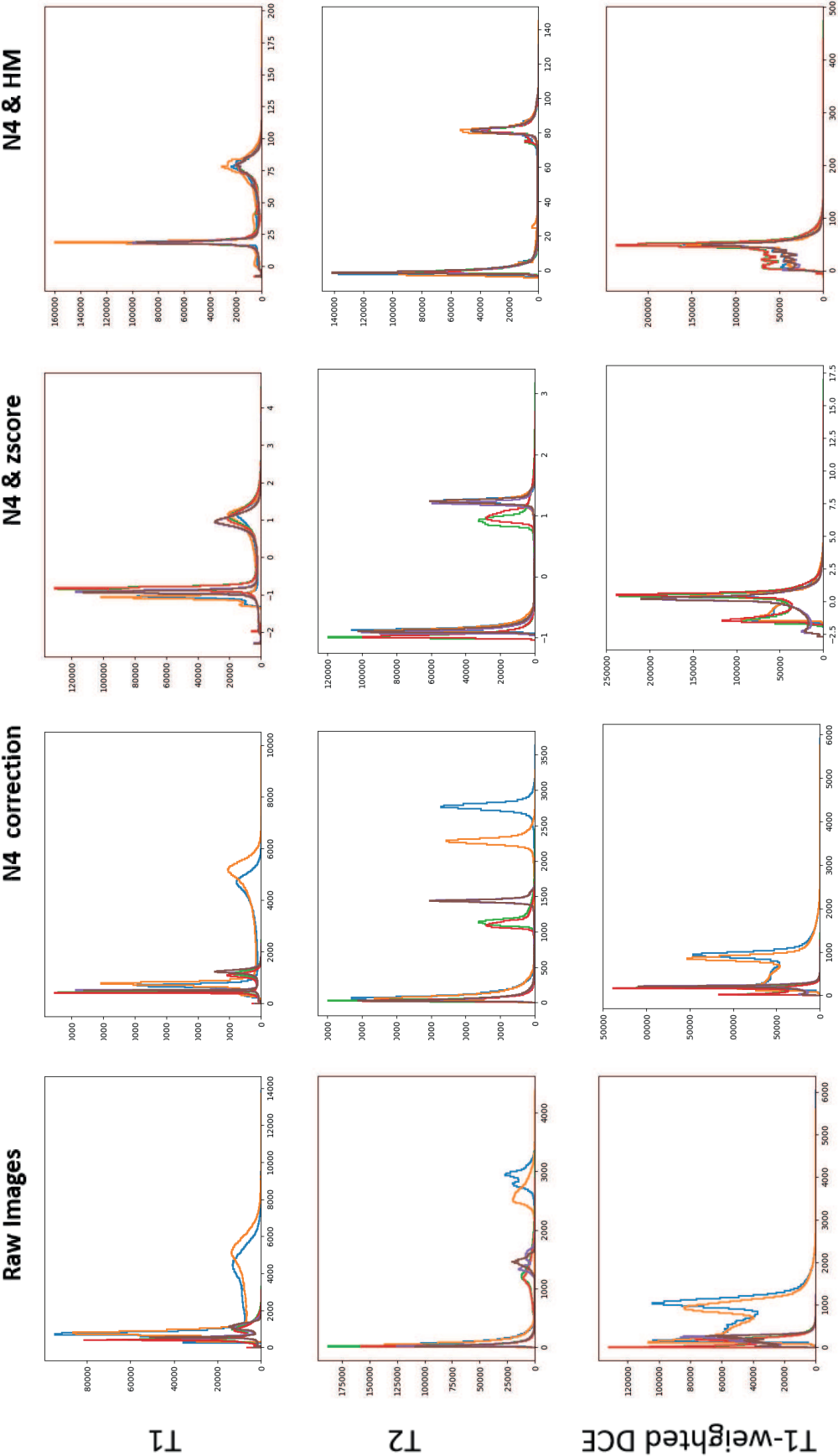


Figure6

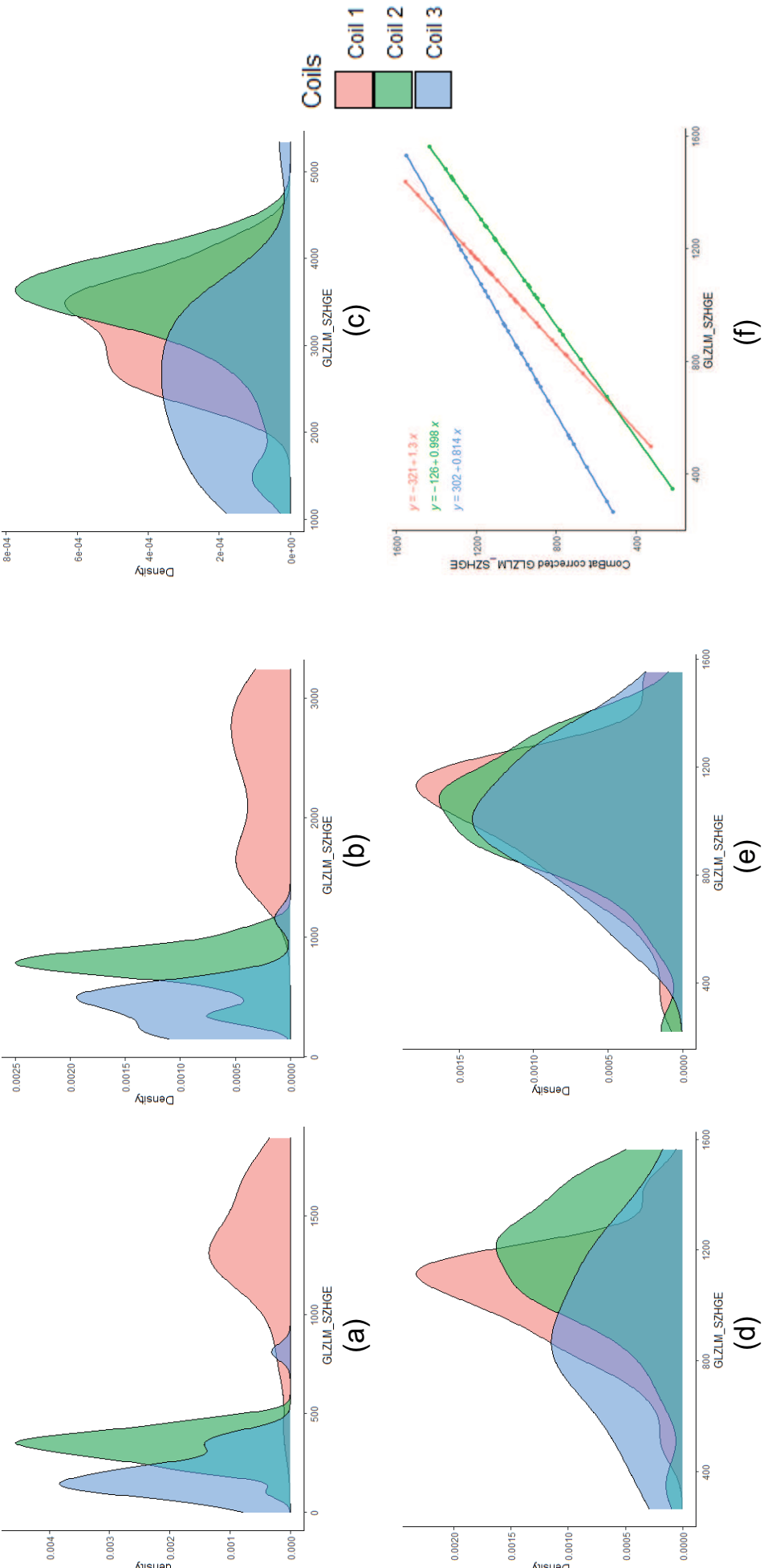


Figure7

[Click here to access/download;Figure;Fig7.eps](#)

