IOP Publishing Phys. Med. Biol. 65 (2020) 015010 (10pp) https://doi.org/10.1088/1361-6560/ab6177

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets

R N Mahon[1,4], M Ghita[2], G D Hugo[3] and E Weiss[1]

[1] Department of Radiation Oncology, Virginia Commonwealth University, Richmond, VA, United States of America
[2] Department of Radiology, Virginia Commonwealth University, Richmond, VA, United States of America
[3] Department of Radiation Oncology, Washington University, St. Louis, MO, United States of America
[4] Author to whom any correspondence should be addressed.

E-mail: mahonrn@mymail.vcu.edu

RECEIVED
14 June 2019

REVISED
3 December 2019

ACCEPTED FOR PUBLICATION
13 December 2019

PUBLISHED
13 January 2020

Keywords: radiomics, non-small cell lung cancer, computed tomography, harmonization

Supplementary material for this article is available online

Abstract

This work seeks to evaluate the combatting batch effect (ComBat) harmonization algorithm's ability to reduce the variation in radiomic features arising from different imaging protocols and independently verify published results. The Gammex computed tomography (CT) electron density phantom and Quasar body phantom were imaged using 32 different chest imaging protocols. 107 radiomic features were extracted from 15 spatially varying spherical contours between 1.5 cm and 3 cm in each of the lung300 density, lung450 density, and wood inserts. The Kolmogorov–Smirnov test was used to determine significant differences in the distribution of the features and the concordance correlation coefficient (CCC) was used to measure the repeatability of the features from each protocol variation class (kVp, pitch, etc) before and after ComBat harmonization. *P*-values were corrected for multiple comparisons using the Benjamini–Hochberg–Yekutieli procedure. Finally, the ComBat algorithm was applied to human subject data using six different thorax imaging protocols with 135 patients. Spherical contours of un-irradiated lung (2 cm) and vertebral bone (1 cm) were used for radiomic feature extraction. ComBat harmonization reduced the percentage of features from significantly different distributions to 0%–2% or preserved 0% across all protocol variations for the lung300, lung450 and wood inserts. For the human subject data, ComBat harmonization reduced the percentage of significantly different features from 0%–59% for bone and 0%–19% for lung to 0% for both. This work verifies previously published results and demonstrates that ComBat harmonization is an effective means to harmonize radiomic features extracted from different imaging protocols to allow comparisons in large multi-institution datasets. Biological variation can be explicitly preserved by providing the ComBat algorithm with clinical or biological variables to protect. ComBat harmonization should be tested for its effect on predictive models.

1. Introduction

Radiomics is an area of active research that seeks to apply computer vision techniques including texture and geometric image feature extraction for use in predictive modelling and machine learning applications.(Aerts *et al* 2014, Fave *et al* 2017, Parekh and Jacobs 2017) Radiomics has been applied to various clinical endpoints and anatomical sites with varying levels of success. The radiomic work flow is comprised of four basic steps (1) image acquisition, (2) target definition, (3) feature extraction, and (4) analysis. While each step of the radiomic process has challenges, the image acquisition is the foundation of the process. CT images are the basis of radiation therapy planning and are an integral part of diagnosis and treatment follow-up. Several CT image acquisition parameters have been shown to effect the extracted radiomic features including contrast enhancement, (He *et al* 2016) slice thickness, (Mackin *et al* 2015, He *et al* 2016, Lu *et al* 2016, Zhao *et al* 2016) reconstruction algorithms, (Mackin *et al* 2015, He *et al* 2016, Kim *et al* 2016, Lu *et al* 2016, Zhao *et al* 2016) voxel size, (Mackin *et al* 2015, Shafiq-Ul-Hassan *et al* 2017) and number of grey levels (Shafiq-Ul-Hassan *et al* 2017). As several imaging

parameters impact the resulting radiomic features, there is a need for standardization and/or harmonization (Yip and Aerts 2016, Zwanenburg *et al* 2016). Standardization, in this context, involves implementing a defined procedure for the radiomic process including (a) using set imaging protocols that produce consistent images for the extraction of radiomic features across different scanners and institutions, and (b) applying a defined methodology for extracting the radiomic features to ensure high fidelity. Harmonization, on the other hand, is reconciling the difference in the radiomic feature values due to changes in imaging protocol or institution after the images and/or radiomic features have been acquired.

In order to reduce the impact of different imaging protocols in studies using radiomics to predict clinical endpoints, many studies will use the same imaging protocol, (Hunter *et al* 2015, Fave *et al* 2017) or controlled protocols (Ger *et al* 2018). Using specified imaging parameters and controlled protocols may help combat the imaging parameter variations in prospective studies, but for retrospective studies, a controlled protocol approach would require either patients to be re-imaged with a certain imaging protocol, which is impractical, or limiting data to only those derived from one imaging protocol. Harmonization of pixel size through resampling (Mackin *et al* 2017) and grey level normalization (Shafiq-Ul-Hassan *et al* 2018) have been applied in retrospective studies, but these techniques can change the pixel or voxel values which may not be desirable in all cases. Isotropic resampling of the voxels, a common pre-processing step in radiomic workflows, requires either up-sampling or down-sampling the image. This can introduce uncertainty due to the interpolation algorithm or reduction of the fine detail in an image, but it can also reduce the impact of different pixel sizes (Mackin *et al* 2017). Grey level normalization involves binning the grey levels in the image prior to feature extraction into a standardized number or width of bins to reduce the impact of noise on the feature values (Shafiq-Ul-Hassan *et al* 2018). Another approach is to apply a *z*-score normalization to the radiomic features for comparison. This technique does not change the image values, but has not been successful in harmonizing all the differences in the features from different imaging parameters (Lu *et al* 2016). The ComBat harmonization method differs from z-score normalization, grey level normalization and resampling in two distinct manners: first, it is not applied to the image itself, and second, it estimates the fraction of total variance across the feature values due to the differences in the imaging protocols before correcting the feature values to remove the batch effect.

Recently, Orlhac *et al* proposed using the ComBat harmonization technique from the field of genetics to correct for batch differences for PET (Orlhac *et al* 2018) and CT (Orlhac *et al* 2019) images. The advantage of the ComBat technique is that the correction is applied directly to the derived radiomic features post extraction as opposed to the image values pre-extraction. This could allow for an easier comparison of features from retrospective or multi-institution data. Our work seeks to verify the findings of Orlhac *et al* (2019) on an independent phantom study and in an independent cohort of non-small cell lung cancer (NSCLC) patients utilizing a wider range of radiomic features.

2. Materials and methods

2.1. Phantom imaging protocols

The phantom setup consists of the Quasar Respiratory Motion Multi-Purpose Body Phantom (ModusQA, London, Ontario) and the Gammex CT Density Phantom (Sun Nuclear, Melbourne, FL). The Quasar phantom contained a solid wood insert on the phantom left, cube insert in the centre, and a 60-degree air wedge insert on the phantom right. The Gammex phantom was placed adjacent to the Quasar phantom with lung300 and lung450 inserts placed on either side of the lowest insert chambers on the inner circle, see figure 1. This location was chosen to align with the Quasar body phantom.

The phantom setup was imaged using 32 different chest imaging protocols on 2 Siemens CT scanners (Erlangen, Germany), the Sensation 64 and the Definition Flash. The protocol variations included different combinations of 6 reconstruction kernels, 3 pitch/bowtie filter combinations, 4 peak kilovoltage (kVp) levels, 3 slice thicknesses, and 2 pitch values with same bowtie filter (pitch only). Only one protocol attribute was varied at a time except for the pitch/bowtie comparison where both the pitch and bowtie filter were varied simultaneously with all other parameters held constant. See table 1 for additional details on the protocol parameter classes investigated, and see supplemental material table 1 (stacks.iop.org/PMB/65/015010/mmedia) for a full description of all protocol parameter comparisons investigated.

2.2. Patient characteristics

Longitudinal follow-up images between three-months and 24-months from 135 patients treated with stereotactic body radiotherapy (SBRT) for primary NSCLC and lung metastases between 2008 and 2018 were retrospectively analysed. The 135 patients were 52.5% female and 47.5% male and ranged in age from 45.1 to 91.4 years at time of treatment. Stage I NSCLC accounted for 86.7% of patients, with stage II comprising an additional 2.2%. The remaining 11.1% were treated for lung metastases. Approximately 25% of patients had all follow-up images acquired with the same imaging protocol.

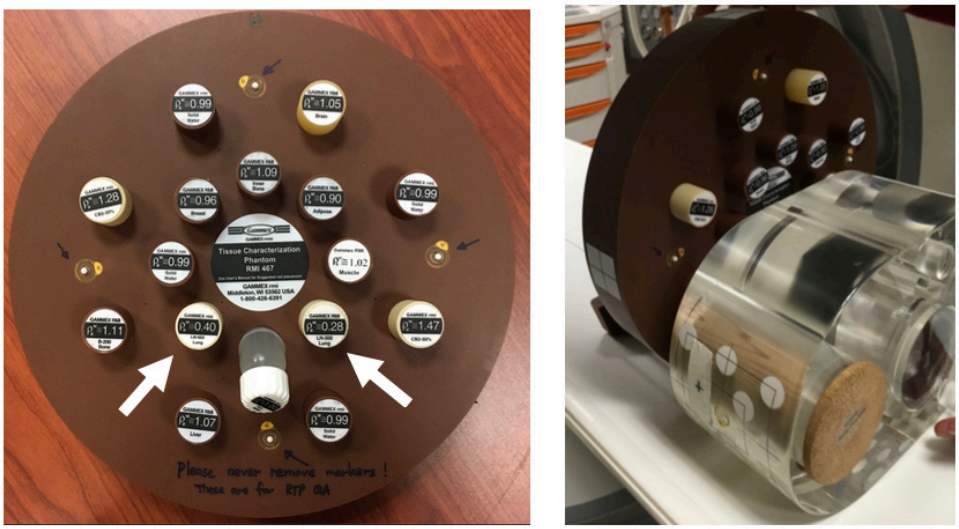

Figure 1. Image of the lung300 (right arrow), lung450 (left arrow) in the Gammex Phantom (left image) and imaging setup for the phantom study (right image).

Table 1. Summary of different imaging protocol classes and variations investigated.

Protocol class	Variations
Slice thickness	2, 3, 5 mm
Pixel width	0.78[a], 0.68[b] mm
kVp	80[a], 100, 120, 140
Pitch only	0.65[a], 1[a]
Pitch/bowtie filter	0.65/Wedge0[a], 2.5/Wedge2[b], 0.6/Wedge3[b]
Reconstruction kernel	B30f, B31f, B40f, B41f, B45f, B70f

[a] Parameter only acquired on Sensation 64 Scanner.
[b] Parameter only acquired on Definition Flash Scanner.

Table 2. Imaging parameters for the six unique patient imaging protocols.

	Parameters						
Protocol	Slice thickness (mm)	Pixel width (mm)	kVp	Bowtie filter	Pitch	Reconstruction kernel	Number of images
Sensation64_A	3	0.58–0.80	100	None	0.65	B40f	12
Sensation64_B	3	0.58–0.80	120	None	0.65	B40f	106
DefinitionFlash_A	3	0.56–0.78	100	Wedge 3	0.6	B41f	44
DefinitionFlash_B	3	0.56–0.78	100	Wedge 3	0.6	B43f	16
DefinitionFlash_C	3	0.56–0.68	120	Wedge 2	0.6	B43f	16
DefinitionFlash_D	3	0.56–0.68	120	Wedge 2	2.5	B41f	9

From the 135 patients identified for this Institutional Review Board approved study, a total of 474 images were available. A subset of 203 images was identified for further analysis by selecting from imaging protocols with at least nine images available for analysis. Six unique imaging protocols were investigated, the details of which can be seen in table 2. Only one image from each protocol was selected per patient. When the patient had more than one image from the same protocol, the earliest time point was used. There was an average interval of 10 ± 6.5 months (range: 3–21 months) between images for patients contributing more than one image from different image protocols.

2.3. Radiomic feature extraction

Spherical regions of interest (ROIs) were delineated on the phantom and patient images using MIM version 6.6 (MIM Software, Cleveland, OH). Within the lung300 and lung450 Gammex inserts, 15 spherical ROIs were drawn with a 1.5 cm diameter. The centre of each sphere was spatially varied while the whole sphere remained within the insert using a one slice offset between contours and allowing overlap. An additional 15 spherical ROIs with a 3 cm diameter were delineated in the wood insert in the Quasar phantom, again with spatially

varying contour centres and using a three slice offset. Each spherical contour from the same image protocol was considered a separate sample instead of imaging the phantoms multiple times.

For the patient images, one spherical ROI was drawn in each of two tissues outside the radiation therapy field: the lung, with a 2 cm diameter, and a vertebra (bone), with a 1 cm diameter. In patients, where multiple tumour lesions were irradiated, regions of the thorax receiving less than 2 Gy of total dose were considered. Only one set of lung and bone contours were delineated per image.

Radiomic texture features were extracted from each ROI using the pyRadiomics (2.2.0) package with python 2.7 (van Griethuysen *et al* 2017). Prior to feature extraction, the images were resampled to the in-plane voxel size using the default B-spline interpolator. No further filters, image pre-processing, or grey level quantization was performed. A total of 107 texture features were extracted from each ROI. The features were from seven categories: shape (14 features), first order or histogram (18 features), the grey level dependence matrix (GLDM) (14 features), the grey level co-occurrence matrix (GLCM) (24 features), the grey level run length matrix (GLRLM) (16 features), the grey level size zone matrix (GLSZM) (16 features), and the neighbourhood grey tone difference matrix (NGTDM) (5 features). A detailed list of the radiomic features extracted can be seen in supplemental material table 2. These radiomic feature categories represent geometrical descriptors as well as basic histogram statistics and higher order texture features which incorporate information regarding the spatial distribution for the grey values within the ROI. A complete mathematical description of the texture features can be found in the pyRadiomics documentation at: https://pyradiomics.readthedocs.io/en/latest/features.html.

2.4. ComBat harmonization

The ComBat harmonization algorithm was originally proposed for the genetics field to address the 'batch effect' seen in microarray analysis. The 'batch effect' refers to non-biological noise, such as operator, time of day the measurements are taken, etc, that affect the assay samples and limits direct comparability. For radiomics studies, the different 'batches' can be thought of as the different image protocols, centres, machines, etc. The ComBat algorithm originally proposed by Johnson and Rabinovic (Johnson *et al* 2007) had the advantage of being effective even with small batch sizes, defined as less than 25 in the aforementioned paper. The ComBat harmonization method is derived from the location and scale (L/S) family of corrections where it is assumed that the error introduced by the batch differences can be corrected by standardizing the means and variances across the batches. In the L/S model, the value Y for feature f from sample j in batch i follows equation (1):

$$Y_{ijf} = a_f + X\beta_f + \gamma_{if} + \delta_{if}\varepsilon_{ijf}. \tag{1}$$

a_f is the overall feature value, X is the design matrix for the sample, β_f is the coefficient matrix for the design matrix, γ_{if} is the additive batch effect, δ_{if} is the multiplicative batch effect and ε_{ijf} is the error, assumed to have a normal distribution with mean 0 and variance σ_f^2. By estimating the additive and multiplicative batch effect parameters for each feature, the corrected feature value, Y^*_{ijf} can be found using equation (2):

$$Y^*_{ijf} = \frac{Y_{ijf} - \widehat{a}_f - X\widehat{\beta}_f - \widehat{\gamma}_{if}}{\delta_{if}} + \widehat{a}_f + X\widehat{\beta}_f \tag{2}$$

where the hat ($\widehat{\ }$) indicates the estimated parameters from the model in equation (1) (Johnson *et al* 2007). The parameter estimates can be determined by applying assumptions about the underlying probability distributions of the additive and multiplicative terms, such as they follow a normal and inverse gamma distribution, respectively. Alternatively, these parameter estimates can be determined empirically using non-parametric assumptions. In addition, known biological variations can be preserved by supplying a model to the ComBat algorithm as the design matrix, X, and associated parameters, $\widehat{\beta}_f$, in equation (1). This work utilized the code developed by Fortin *et al* (2017, 2018), which can be found at https://github.com/Jfortin1/ComBatHarmonization, using R (3.5.1) and R studio (1.1.456, R Studios Inc., Boston, MA) with the non-parametric settings and without a biological model.

2.5. Statistical analysis

For the phantom portion of this work, comparisons in protocol differences were made on a class-wise basis where all image parameters were held at a 'standard' protocol level except the class being evaluated. The 'standard' protocol was acquired at 100 kVp, with a slice thickness of 3 mm, and was reconstructed using the B41f kernel for both machines. As the same bowtie filters were not available on both machines, the 'standard' protocol included a pitch of 0.65 and no bowtie filter (Wedge0) on the Sensation 64 and a pitch of 0.6 and the Wedge3 bowtie filter on the Definition Flash. When performing the ComBat harmonization, each parameter variation was considered a different batch. For example, when harmonizing the reconstruction parameter class, there were six different batches harmonized at once.

Table 3. Percentage of 107 radiomic features derived from significantly different distributions before and after harmonization for all protocol comparisons.

	Lung300		Lung450		Wood		
	Before	After	Before	After	Before	After	Number of comparisons per contour
Pitch/Bowtie Filter	64%–72%	0%	57%–66%	0%	11%–50%	0%–1%	3
kVp	0%–83%	0%	31%–78%	0%	0%–68%	0%	9
Pitch	69%–75%	0%	66%–68%	0%	45%–50%	0%	2
Reconstruction Kernel	7%–87%	0%	7%–85%	0%–2%	0%–78%	0%	30
Slice Thickness	43%–67%	0%–1%	40%–71%	0%	4%–66%	0%	6

For the patient data, comparisons were also made on a class-wise basis. However, all imaging protocols with the desired parameter variation were considered together to mimic a closer to real life situation and test the ComBat harmonization algorithm's ability to correct for multiple imaging acquisition parameter variations at once. For example, two of the six imaging protocols were acquired at 100 kVp, while the other four were acquired at 120 kVp. All images acquired with 100 kVp would be compared to all images acquired with 120 kVp regardless of which other parameters were varied. When performing the ComBat harmonization, all images were harmonized together resulting in six different batches.

The 2-sample Kolmogorov–Smirnov test was used to determine if there was a significant difference in the distribution of the features from each protocol variation class (kVp, pitch, etc) both before and after ComBat harmonization. The 2-sample Kolmogorov–Smirnov test is a non-parametric test of whether two samples are derived from the same underlying probability distribution. Repeatability of texture features was assessed for all protocol variations using the concordance correlation coefficient (CCC) as described by Lin (1989). The CCC measures the correlation between two paired measurements by calculating the deviation from one-to-one correlation. For all comparisons, a threshold of 0.9 was used to determine whether or not a feature was repeatable, as suggested by McBride (2005). All p-values were corrected for multiple comparisons using the Benjamini-Hochberg-Yekutieli procedure (Benjamini and Hochberg 1995, Benjamini and Yekutieli 2001) using the built in p.adjust function in R.

3. Results

3.1. Phantom results

The ComBat harmonization procedure reduced the percentage of features having significantly different underlying distributions in a variety of situations. For the 150 different image protocol comparisons there was a large reduction, or a constant 0%, in the percentage of features that were from significantly different distributions after harmonization, see table 3. None of the image protocol comparisons resulted in an increase in the number of features from significantly different distributions. Following harmonization, there were five comparisons that still exhibited a significant difference in the distributions: slice thickness 3 mm versus 5 mm for the lung300 insert from the Definition Flash, the reconstruction filter comparisons of B30f to B31f, B30f to B40f, and B30f to B41f for the lung450 insert on the Sensation 64, and the filter comparison of Wedge0 to Wedge2 for the wood insert with one image being acquired on each machine. An example of the distribution of the long run emphasis texture features for the different pitch/bowtie filter combinations before and after ComBat harmonization for the lung300 insert, figure 2, demonstrates a successful harmonization. Examples of the protocol comparisons where the harmonization failed can be seen in supplemental material figures 1 and 2.

The repeatability of features between the 150 different comparisons as measured by the CCC followed a similar pattern as the analysis of the significantly different distributions. For all comparisons, the number of features that had a CCC score greater than 0.9 remained the same or increased following ComBat harmonization. For the case of the Pitch/Bowtie Filter comparisons, the differences in the images resulted in no features being repeatable before or after harmonization. However, on average, an additional 23 features became repeatable following harmonization across all image comparisons for all features.

3.2. Patient results

The six different image protocols resulted in seven protocol variation class comparisons: kVp 100 versus 120; reconstruction kernels B40f versus B41f, B40f versus B43f, and B41f versus B43f; and pitch/bowtie filter combinations of 0.65 pitch /Wedge0 versus 0.6 pitch/Wedge3, 0.65 pitch/Wedge0 versus 2.5 pitch/Wedge2, and 0.6 pitch/Wedge3 versus 2.5 pitch/Wedge2. A large range of variation in the average relative feature value difference between protocol comparisons before ComBat harmonization can be seen for both the bone (figure 3) and lung (figure 4). The shape features show no variation before harmonization as expected since a uniform shape

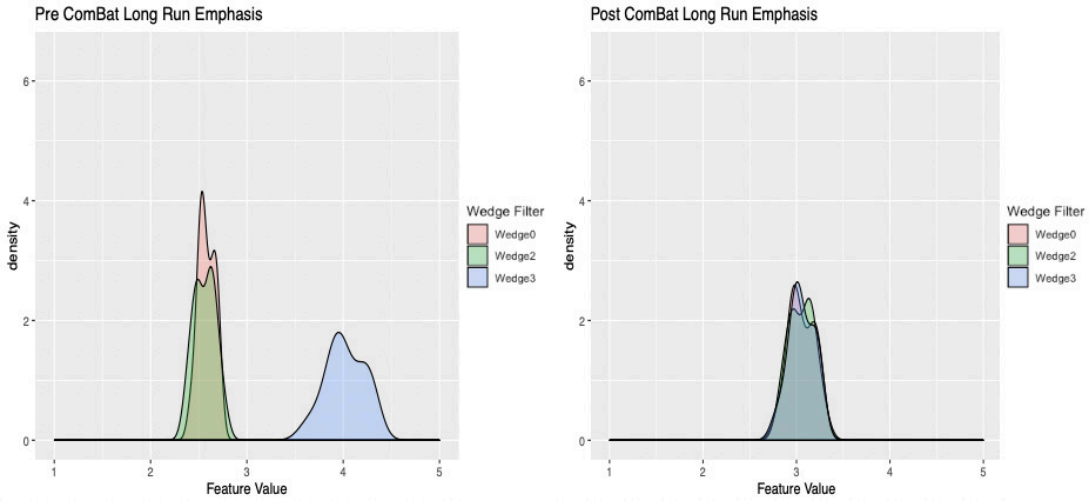

Figure 2. Example of the probability density curves for the long run emphasis feature from the GLRLM before ComBat harmonization (left image) and after ComBat harmonization (right image) for the lung300 phantom contour. Each fill colour represents a different pitch/ bowtie filter combination: 0.65pitch/Wedge0 (Wedge0—red); pitch 2.5/Wedge2 (Wedge2—green); and pitch 0.6/Wedge3 (Wedge3—blue).

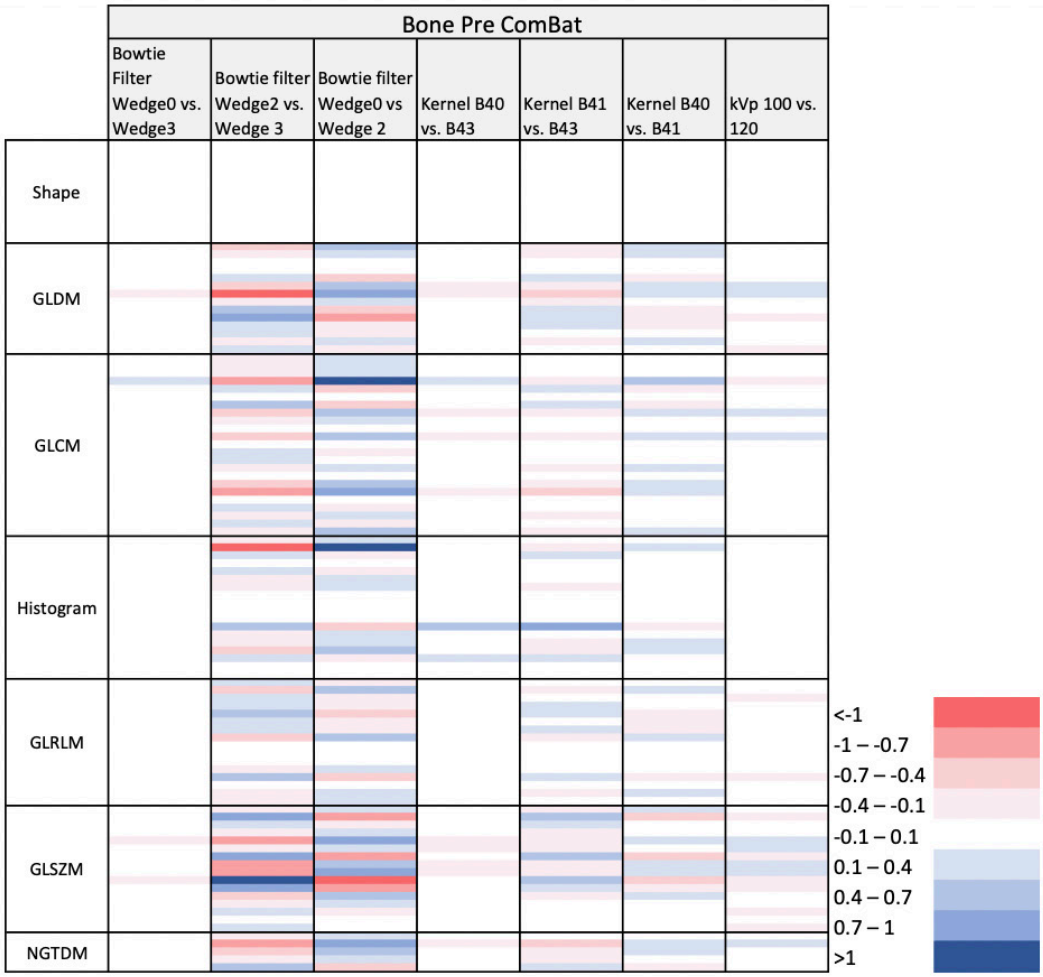

Figure 3. Heat map of relative difference between average texture feature values for the seven protocol variation comparisons for bone contour before ComBat harmonization. Features include the shape, GLDM, GLCM, histogram, GLRLM, GLSZM, and NGTDM features.

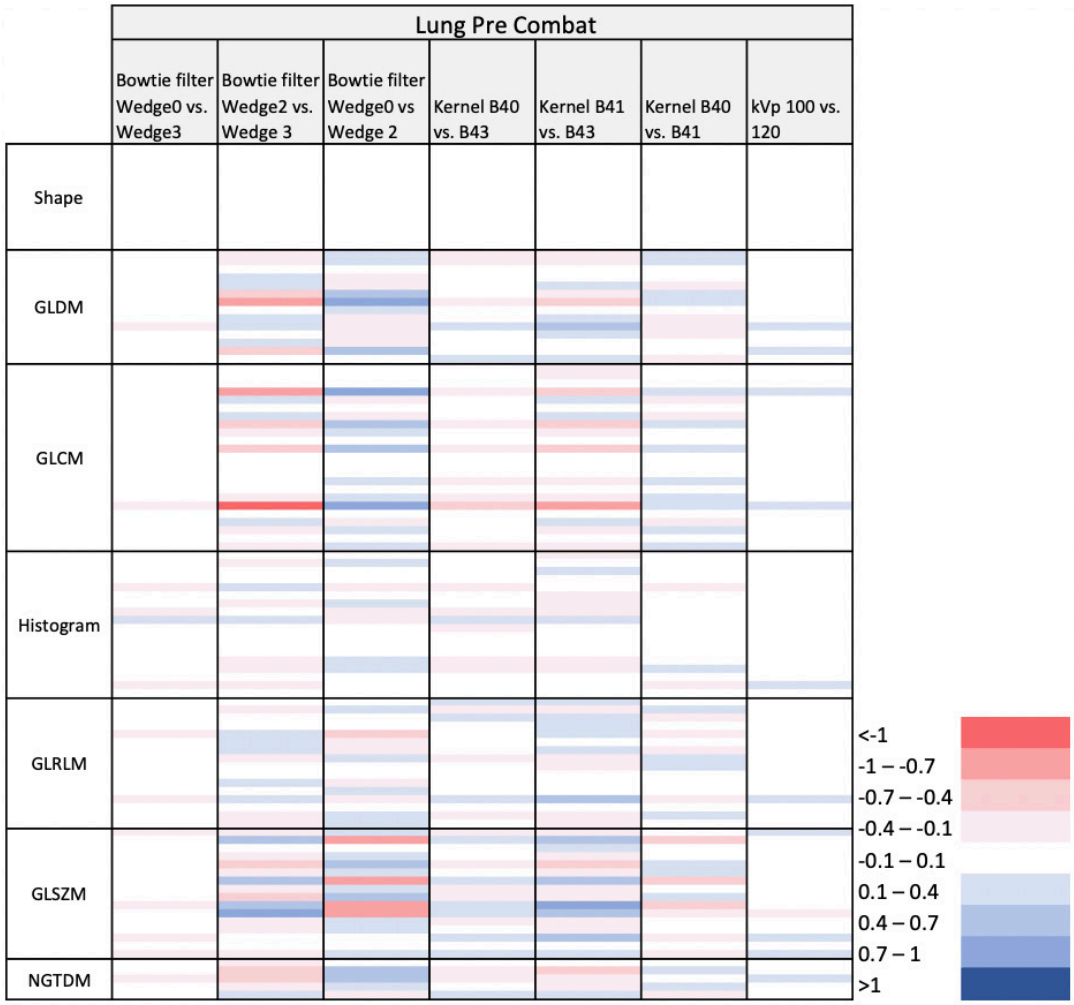

Figure 4. Heat map of relative difference between average texture feature values for the seven protocol variation comparisons for lung contour before ComBat harmonization. Features include the shape, GLDM, GLCM, histogram, GLRLM, GLSZM, and NGTDM features.

and size contour was created in each image. Following harmonization, the average relative difference between all radiomic features was between −0.1 and 0.1. ComBat harmonization was able to reduce the percentage of radiomic features from significantly different distributions to 0% in all cases, see table 4. The distribution of the long run emphasis example feature for the patient data before and after harmonization can be seen in figure 5.

4. Discussion

This work sought to evaluate the ComBat harmonization algorithm on data from our institution and verify the finding of Orlhac *et al* (2019) regarding the use of ComBat harmonization in CT radiomics studies using additional features and applying it to a larger patient cohort. This work was able to demonstrate the ComBat harmonization algorithm's ability to correct radiomic features from a variety of different imaging protocols in both phantom and patient datasets. The ComBat method has the advantage of correcting the radiomic features post extraction and not modifying the original images making it potentially easier to share data in multi-centre studies as the image data themselves would not have to be shared. In addition, it provides a method of correcting feature values from retrospective, multi-centre, and/or longitudinal studies so they are comparable and potentially able to be used in predictive modelling.

When compared to the Orlhac *et al* (2019) paper, the phantom portion of our work investigated similar differences in imaging protocol: reconstruction kernels, slice thickness, and spiral pitch factor. While Orlhac *et al* allowed the effective milliamperage, manufacturer, and voxel size to also vary, this work allowed the kVp to vary. Whereas Orlhac *et al* used a 10 HU binning of the grey level values prior to feature extraction in the phantom data, we did not apply grey level quantization. Binning the grey level values is a common practice in radiomics as it can reduce the impact of noise on the feature values at the expense of smoothing over finer texture patterns. However, there is not a consensus on the width of the bin to use (Zwanenburg *et al* 2016). By not quantizing the

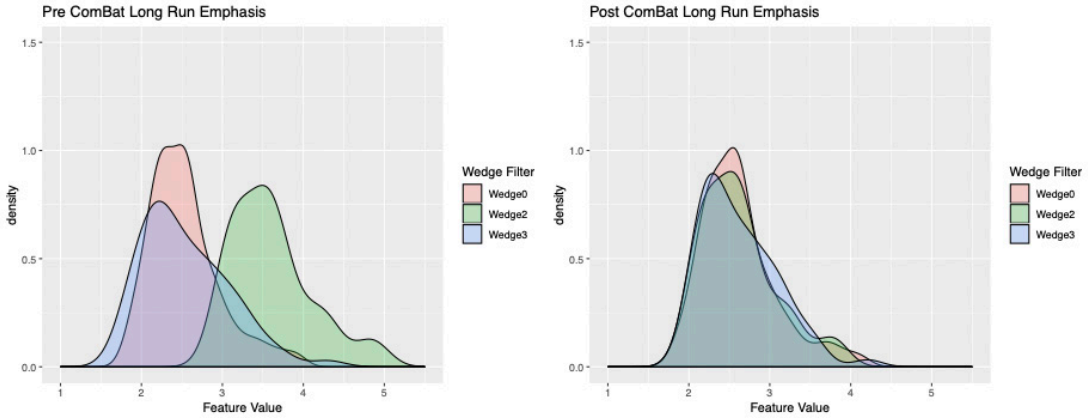

Figure 5. Example of the probability density curves for the long run emphasis feature from the GLRLM before ComBat harmonization (left image) and after ComBat harmonization (right image) for the bone contour. Each fill colour represents a different pitch/ bowtie filter combination: pitch 0.65/Wedge0 (Wedge0—red); pitch 2.5/Wedge2 (Wedge2—green); and pitch 0.6/ Wedge3 (Wedge3—blue).

Table 4. Summary of percentage of radiomic features derived from significantly different distributions for the patient images before and after ComBat harmonization.

	Bone		Lung	
	Before	After	Before	After
Pitch/bowtie filter	0%–62%	0%	0%–19%	0%
kVp	2%	0%	0%	0%
Reconstruction kernel	0%	0%	0%	0%

grey level, this work can be seen as testing the technique on a 'more challenging' situation with regards to image noise impacting the feature values. This work chose the lung300, lung450 and wood phantom for analysis as they have an inherent texture pattern. Both the lung300 and lung450 are foam like inserts with different size and distribution of air bubbles to mimic the density of lungs. The wood insert has a natural wood grain texture with more directional dependence than the foam inserts. These inserts are similar to the cork and sycamore wood inserts used in the Orlhac *et al* paper. In addition, this work expanded the number of texture features investigated from 40, for the phantom data, with the LIEFx freeware (www.lifexsoft.org) used by Orlhac *et al*, to 107 texture features extracted using the pyRadiomics (van Griethuysen *et al* 2017) software. The additional features were predominately from the addition of the GLDM features and shape features. Both the pyRadiomics and LIFEx software programs calculate features from the first order histogram, GLCM, GLRLM, GLSZM, and NGTDM. The patient data in this work was expanded to include 135 patients and six different protocols compared to the 74 patients and six protocols analysed by Orlhac *et al*. In addition, the same 107 features were analysed on the patient data. Orlhac *et al* examined 10 to 89 features extracted from two different in-house software packages on their patient data. The results of this work are in good agreement with the findings of the aforementioned paper. Mainly, the ComBat harmonization was able to reduce or remove the batch effect for both the phantom and patient data sets in both studies. The ability of the ComBat harmonization to perform well in different imaging scenarios demonstrates the flexibility of the procedure.

The ComBat harmonization technique is based on estimating the statistical variance present among the different batches. This procedure has the same goal as a normalization procedure but with subtle differences. In a traditional normalization procedure, like dividing by a reference value or *z*-score normalization, metrics such as the overall mean and standard deviation are used to determine the deviation from 'normal'. In the Combat algorithm, only a fraction of the total variance is removed. The algorithm estimates the fraction of the total variance across the different batches that is due only to the differences in the imaging protocols for each feature independently without the use of a reference. Normalization procedures do not normally target a single source of variation, but adjust the values on overall variation, while the ComBat algorithm seeks to preserve real difference in the data. In addition, the ComBat harmonization estimates two contribution modes of the effect, the additive and the multiplicative effects. Most normalization procedures do not estimate multiple modes. As harmonization could lead to some of the variation explained by true clinical or biological differences in the data to be inadvertently corrected, the ComBat algorithm does allow for the option of preserving known or suspected biological variation explicitly by providing the algorithm with a model that uses the desired biological features as covariates.

A few general trends in image acquisition parameters could be seen in the pre harmonization percentage of features derived from significantly different distributions. The number of features from significantly different distributions appeared directly proportional to the difference in acquisition kVp for comparisons between 100 kVp and 140kVp. In addition, all comparisons with 80 kVp had a larger number of features from significantly different distributions than comparisons among the 100–140 kVp range. The comparisons with 80 kVp still showed an increase in the number of features from significantly different distributions with increasing gap in acquisition kVp. A similar directly proportional trend was seen between the number of significantly different features and the difference in slice thickness. However, the image protocols acquired on the Sensation 64 tended to have a higher number of significantly different features than the Definition Flash images when evaluating the same slice thickness comparisons, suggesting a difference between the machines in addition to the imaging parameters. Finally, the reconstruction kernels B30f, B31f, B40f, and B41f all had a comparable number of significant features. There was approximately a 15%–30% increase in the percentage of features from significantly different distributions when any reconstruction kernel was compared to the B45f and B70f kernels. As the number in the name of the reconstruction kernel increases, the sharpness of the kernel also increases to emphasize high spatial frequency features and hence less smoothing occurs. The B30f, B40f and B70f are base kernels with the B31f and B41f differing by having a finer grain noise and milder edge enhancement than their corresponding bases. The B45f kernel has a sharpness half way in between B40f and B50f. The 'B' in the kernel name represents a body kernel while the 'f' represents a fast implementation of the kernel. Given the kernel sharpness, it is understandable that both the B45f and B70f reconstructed images appeared noisier than the other tested reconstruction kernels. This increase in noise could cause the observed higher percentage of features from significantly different distributions which may be mitigated by applying grey level quantization.

Five image protocol comparisons were not harmonized following the ComBat harmonization procedure. In four of the five instances, cluster prominence from the GLCM was unsuccessfully harmonized. In the fifth instance, kurtosis from the intensity features was not successfully harmonized. Cluster prominence is a measure of uniformity and proximity as it relates to perception in an image (Conners *et al* 1984). Kurtosis is a measure of the sharpness of the peak in the distribution of pixel values. These features could have still shown a significant difference after harmonization in part due to a multi-modal distribution of values. The Kolmogorov-Smirnov test measures the maximum vertical distance between two probability density curves. The existence of a shift between the mean peaks can increase the maximum vertical distance causing the Kolmogorov–Smirnov test to be more sensitive to shifts in the mean (Berger and Zhou 2005). With the cluster prominence and kurtosis features, the distribution before harmonization had one dominant peak, with additional smaller peaks further away. For the kurtosis feature, the additional peaks were relatively close to the dominate peak, while the cluster prominence was spread over a larger area. Following harmonization, for both features, the peaks moved closer together, but a widening shift can be seen between the centres of the dominant peaks. Supplemental material figures 1 and 2 illustrate this resulting shift in the peaks that may contribute to their significant difference as measured by the Kolmogorov–Smirnov test.

An alternative method to the ComBat approach, a post extraction standardization method proposed by Andrearczyk *et al* (2019) uses a two layer multi-layer perceptron (MLP) network to harmonize features. The Andrearczyk *et al* phantom study showed the MLP based harmonization was able to increase the inter-class correlation for the features from 0.63 to 0.78 averaged across all features. However, this process requires an adequate training set to determine the network weights. In comparison, the ComBat harmonization method does not need extensive training. In both methods, the weights for MPL, and the correction parameters for ComBat, are dependent on the data provided to the algorithm. In theory, the MPL method applied to a new data set could make use of transfer learning as a starting point for the weights, but training with the new data set may still be needed, which could be time consuming. The ComBat harmonization is applied to each new set of data to determine the correction parameters with results in a matter of seconds.

While this work demonstrated the effectiveness of the ComBat harmonization on the radiomic features, it did not investigate the effect of adding a biological model to the ComBat algorithm. The biological model optionally allows the user to explicitly prevent variation due to clinical or other biological factors from accidently being removed in addition to the batch effects. Providing this model with biological parameters allows the iterative algorithm to estimate the fraction of the total variation due to the parameters included in the biological model and protects this variation. The algorithm then derives the parameter estimates for the batch effect on the remaining non protected variation. The use of a biological model should be investigated fully. In addition, the effect of using harmonized parameters on predictive models for clinical endpoints should be investigated as a future work.

5. Conclusion

This work demonstrates the ability of the ComBat harmonization algorithm to make radiomic features from different imaging protocols comparable and validates previously published results. In addition, this

harmonization method has the ability to increase the repeatability of texture features. ComBat harmonization is easy to implement and does not affect the original imaging data making it ideal for retrospective, longitudinal, or multi-institutional studies. The effect of ComBat harmonization on imaging-feature based predictive models needs to be investigated.

Acknowledgment

This work was supported in part by a research grant from Varian Medical Systems.

ORCID iDs

R N Mahon https://orcid.org/0000-0003-0466-2282

References

Aerts H J W L *et al* 2014 Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach *Nat. Commun.* **5** 4006

Andrearczyk V, Depeursinge A and Mueller H 2019 Learning cross-protocol radiomics and deep feature standardization from CT images of texture phantoms *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications* ed P R Bak and P-H Chen (Bellingham, WA: SPIE Optical Engineering Press) p 17 (www.spiedigitallibrary.org/conference-proceedings-of-spie/10954/2512683/Learning-cross-protocol-radiomics-and-deep-feature-standardization-from-CT/10.1117/12.2512683.full)

Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc.* **57** 289–300

Benjamini Y and Yekutieli D 2001 The control of the false discovery rate in multiple multiple testing under dependency *Ann. Stat.* **29** 1165–88

Berger V W and Zhou Y 2005 Kolmogorov–Smirnov tests *Encyclopedia of Statistics in Behavioral Science* (Chichester: Wiley) pp 1–5

Conners R W, Trivedi M M and Harlow C A 1984 Segmentation of a high-resolution urban scene using texture operators *Comput. Vis. Graph. Image Process.* **25** 273–310

Fave X *et al* 2017 Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer *Sci. Rep.* **7** 588

Fortin J-P *et al* 2018 Harmonization of cortical thickness measurements across scanners and sites *NeuroImage* **167** 104–20

Fortin J-P *et al* 2017 Harmonization of multi-site diffusion tensor imaging data *NeuroImage* **161** 149–70

Ger R B *et al* 2018 Comprehensive investigation on controlling for CT imaging variabilities in radiomics studies *Sci. Rep.* **8** 1–14

He L, Huang Y, Ma Z, Liang C, Liang C and Liu Z 2016 Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule *Sci. Rep.* **6** 1–10

Hunter L A *et al* 2015 NSCLC tumor shrinkage prediction using quantitative image features *Comput. Med. Imaging Graph.* **49** 29–36

Johnson W E, Li C and Rabinovic A 2007 Adjusting batch effects in microarray expression data using empirical Bayes methods *Biostatistics* **8** 118–27

Kim H, Park C M, Lee M, Park S J, Song Y S, Lee J H, Hwang E J and Goo J M 2016 Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability *PLoS One* **11** 1–11

Lin L I 1989 A concordance correlation-coefficient to evaluate reproducibility *Biometrics* **45** 255–68

Lu L, Ehmke R C, Schwartz L H and Zhao B 2016 Assessing agreement between radiomic features computed for multiple CT imaging settings *PLoS One* **11** e0166550

Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, Rodriguez-Rivera E, Dodge C, Jones A K and Court L 2015 Measuring computed tomography scanner variability of radiomics features *Invest. Radiol.* **50** 1–9

Mackin D, Fave X, Zhang L, Yang J, Jones A K, Ng C S and Court L 2017 Harmonizing the pixel size in retrospective computed tomography radiomics studies *PLoS One* **12** 1–17

McBride G 2005 A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient *NIWA Client Rep.* **HAM2005-06** 14

Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V and Buvat I 2018 A postreconstruction harmonization method for multicenter radiomic studies in PET *J. Nucl. Med.* **59** 1321–8

Orlhac F, Frouin F, Nioche C, Ayache N and Buvat I 2019 Validation of a method to compensate multicenter effects affecting CT radiomics *Radiology* **291** 53–9

Parekh V S and Jacobs M A 2017 Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI *npj Breast Cancer* **3** 43

Shafiq-Ul-Hassan M *et al* 2017 Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels *Med. Phys.* **44** 1050–62

Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R and Moros E 2018 Voxel size and gray level normalization of CT radiomic features in lung cancer *Sci. Rep.* **8** 1–9

van Griethuysen J J M, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan R G H, Fillion-Robin J-C, Pieper S and Aerts H J W L 2017 Computational radiomics system to decode the radiographic phenotype *Cancer Res.* **77** e104–7

Yip S S F and Aerts H J W L 2016 Applications and limitations of radiomics *Phys. Med. Biol.* **61** R150–66

Zhao B, Tan Y, Tsai W Y, Qi J, Xie C, Lu L and Schwartz L H 2016 Reproducibility of radiomics for deciphering tumor phenotype with imaging *Sci. Rep.* **6** 1–7

Zwanenburg A, Leger S, Vallières M and Löck S 2016 Image biomarker standardisation initiative (arXiv:1612.07003)