Evaluation of conventional and deep learning based image harmonization methods in radiomics studies

F Tixier, Vincent Jaouen, C Hognon, O Gallinato, T Colin, D Visvikis

IOP Publishing *Phys. Med. Biol.* **00** (2021) 000000

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

CrossMark

Q1 Q2

RECEIVED
Q3 17 June 2021

REVISED
9 November 2021

Q4 ACCEPTED FOR PUBLICATION
15 November 2021

PUBLISHED
DD MM 2021

PAPER

Evaluation of conventional and deep learning based image harmonization methods in radiomics studies

F Tixier[1,2], **V Jaouen**[1,3], **C Hognon**[1,4], **O Gallinato**[4], **T Colin**[4] and **D Visvikis**[1]

[1] LaTIM UMR 1101, INSERM, Brest, France
[2] Radiation Therapy Department, Brest University Hospital, Brest, France
[3] IMT Atlantique, Brest, France
[4] Sophia Genetics, Cité de la Photonique, Pessac, France

E-mail: florent.tixier@gmail.com and florent.tixier@univ-brest.fr

Keywords: radiomics, harmonization, machine learning

Abstract

Objective. To evaluate the impact of image harmonization on outcome prediction models using radiomics. *Approach.* 234 patients from the Brain Tumor Image Segmentation Benchmark (BRATS) dataset with T1 MRI were enrolled in this study. Images were harmonized through a reference image using histogram matching (H_{HM}) and a generative adversarial network (GAN)-based method (H_{GAN}). 88 radiomics features were extracted on H_{HM}, H_{GAN} and original (H_{NONE}) images. Wilcoxon paired test was used to identify features significantly impacted by the harmonization protocol used. Radiomic prediction models were built using feature selection with the Least Absolute Shrinkage and Selection Operator (LASSO) and Kaplan–Meier analysis. *Main results.* More than 50% of the features (49/88) were statistically modified by the harmonization with H_{HM} and 55 with H_{GAN} (adjusted p-value < 0.05). The contribution of histogram and texture features selected by the LASSO, in comparison to shape features that were not impacted by harmonization, was higher in harmonized datasets (47% for H_{none}, 62% for H_{HM} and 71% for H_{GAN}). Both image-based harmonization methods allowed to split patients into two groups with significantly different survival ($p<0.05$). With the H_{GAN} images, we were also able to build and validate a model using only features impacted by the harmonization (median survivals of 189 versus 437 days, $p = 0.006$) *Significance.* Data harmonization in a multi-institutional cohort allows to recover the predictive value of some radiomics features that was lost due to differences in the image properties across centers. In terms of ability to build survival prediction models in the BRATS dataset, the loss of power from impacted histogram and heterogeneity features was compensated by the selection of additional shape features. The harmonization using a GAN-based approach outperformed the histogram matching technique, supporting the interest for the development of new advanced harmonization techniques for radiomic analysis purposes.

Introduction

The extraction of biomarkers from medical images, commonly known today as radiomics, has been shown in numerous studies to be a promising tool in oncology for patient management including diagnosis, follow-up, outcome prediction and therapy response (El Naqa *et al* 2009, Gillies *et al* 2015, Hatt *et al* 2017). Radiomics allow to quantify different image characteristics using intensity, textural and shape descriptors that provide useful information to characterize multiple diseases (Lambin *et al* 2012, Zwanenburg *et al* 2016). Then, radiomics features can be combined using machine learning approaches to build predictive and/or prognostic models (Parmar *et al* 2015, Desseroit *et al* 2016, Leger *et al* 2017, Lucia *et al* 2017).

One of the main limitations of these models comes from the reproducibility of the radiomics features impacted by the acquisition and reconstruction protocols, making it difficult to fully exploit the derived models

in a multi-institutional setting (Um *et al* 2019, Da-ano *et al* 2020a). This is particularly true for radiomics on MR images where voxels intensity does not reflect physical characteristics such as electron density in computed tomography (CT) or glucose metabolism in 18F-FDG positron emission tomography images (Kumar *et al* 2012).

For this reason, several strategies have been proposed in order to normalize radiomics features extracted from images acquired with different protocols (Da-ano *et al* 2020a). These strategies can be divided into two main categories consisting in either standardizing radiomics features after their extraction (Orlhac *et al* 2018, Da-ano *et al* 2020b) or image harmonization prior to feature extraction (Hognon *et al* 2019, Um *et al* 2019). The first option is the solution of choice when images are not available (for instance in multi-institutional studies where the transfer of full images can be restricted due to protection of personal health information data). Alternatively, when images are available, working in the image space offers several advantages over the extracted features harmonization strategy. Such advantages include the possibility of single patient data harmonization, otherwise impossible in the case of radiomics features normalization which requires the availability of a center specific cohort for the construction of a harmonization model. In addition, harmonization in the image space allows other steps beyond the radiomics feature extraction to benefit from of harmonization such as for example in the necessary segmentation step which could be biased by the variability of the image quality.

A very limited number of studies have made use of image harmonization to date within the context of radiomics analysis, showing that histogram matching can contribute to reducing radiomics feature variability (Um *et al* 2019). However, the impact on the predictive power of radiomics features obtained from harmonized images remains unclear. In addition, new harmonization methods based on generative adversarial networks (GAN) seem very promising (Goodfellow *et al* 2014, Modanwal *et al* 2020, Zhong *et al* 2020) but their interest in radiomics studies have not been fully investigated yet.

For this reason, this study aimed at comparing predictive models built with radiomics features extracted from images after harmonization using two image-based harmonization techniques, namely histogram matching harmonization and GAN based harmonization) on the Brain Tumor Image Segmentation Benchmark (BRATS) dataset, which is a multi-institutional dataset of brain tumor images (Menze *et al* 2015, Bakas *et al* 2017, 2018).

Patients and methods

Patients

234 patients from the BRATS 2020 dataset (Menze *et al* 2015, Bakas *et al* 2017, 2018) with the available overall survival (OS) information were included in this study. The cohort was randomly divided into a training (187 patients) and a testing (47 patients) dataset. Patients were acquired with different clinical protocols and various scanners from 19 different institutions. This dataset is publicly available through the annual Medical Image Computing and Computer Assisted Intervention (MICCAI) Society brain tumor segmentation challenge and consequently ethical committee approval was not required for the study.

Images and segmentation

All the patients from this dataset had skull-stripped and co-registered T1, post-contrast T1-weighted, T2-weighted and T2 Fluid Attenuated Inversion Recovery images interpolated to a resolution of 1 mm3. Image segmentations used are the ones that are provided in the BRATS dataset, corresponding to manual delineation of the tumors using the same annotation protocol by experienced neuro-radiologists including the Gadolinium (Gd)-enhancing tumor, the peritumoral edema and the necrotic and non-enhancing tumor core (Menze *et al* 2015). In this paper we focused on the analysis of tumor core (Gd-enhancing tumor + necrotic and non-enhancing tumor core) in T1 images.

Image harmonization approaches

We studied two different image harmonization strategies based on (1) histogram matching (HM) (Gonzalez and Woods 2006) and (2) on a previously proposed two-step image harmonization method based on cycleGAN and pix2pix image-to-image translation GAN (Hognon *et al* 2019). The rationale behind the two-step learning process is to reduce overfitting to the target image by first performing a data augmentation step using a short cycleGAN training (20 epochs) between the source and target domains followed by a longer pix2pix training between the source and augmented dataset. Both HM and the GAN-based approach require the identification of a target reference image towards which all source images are matched. The same reference was used for both approaches (reference named STD_2013_29 in the BRATS dataset). HM was implemented in C++ using the ITK library using 256 histogram levels and seven control points. The neural network model was implemented in python using the Keras framework. The generator was a 7-layer U-Net generative architecture starting at resolution 256 × 256 with dropout rate of 0.2 and batch normalization.

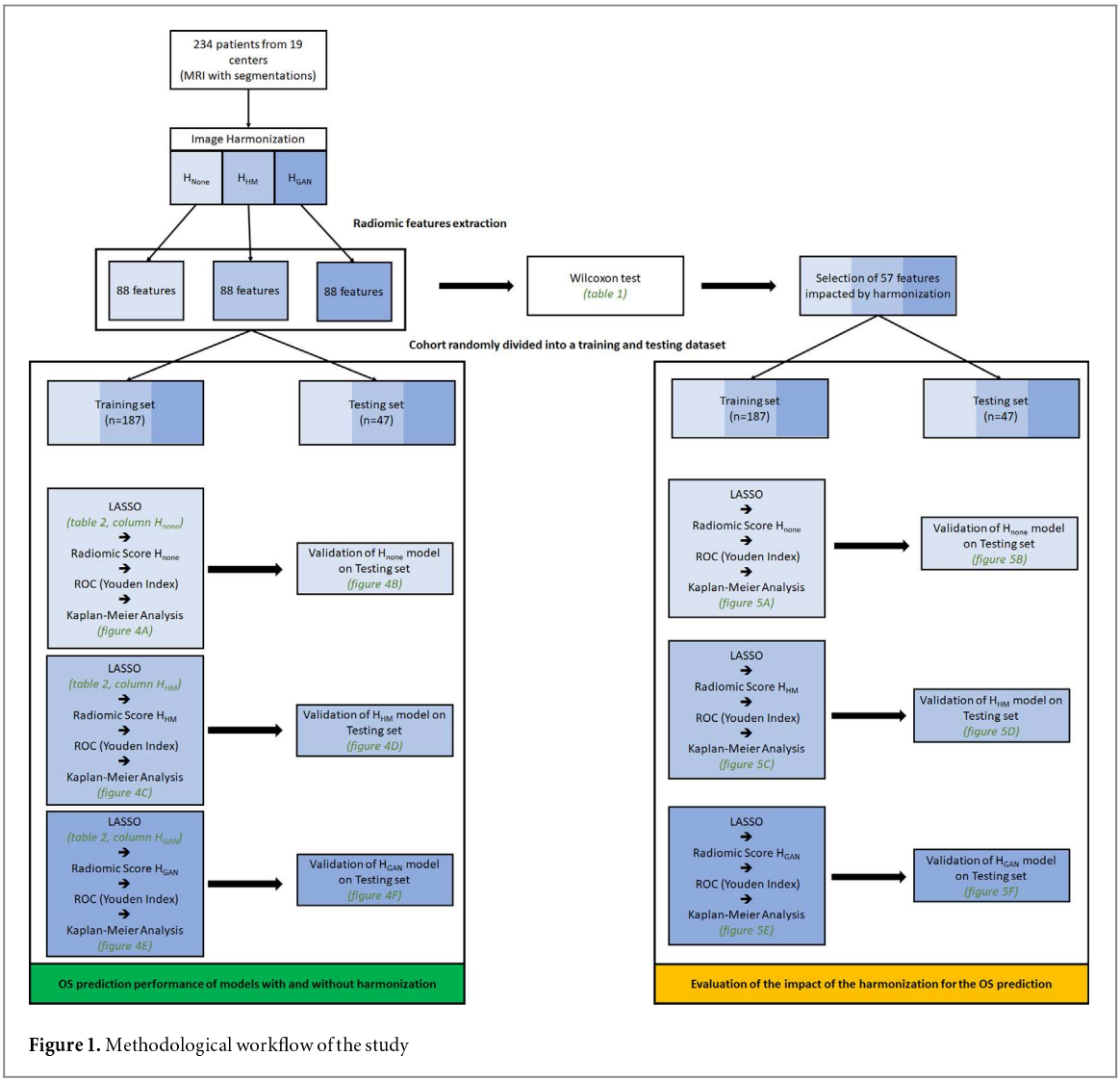

Figure 1. Methodological workflow of the study

Radiomics features extraction

Radiomics features were computed directly from images in the BRATS dataset without harmonization (H_{none}) and from harmonized images after histogram matching harmonization (H_{HM}) and GAN based harmonization (H_{GAN}).

A set of 88 radiomics features were extracted from the most common features categories using an in-house IBSI compliant software (Zwanenburg *et al* 2016): 11 intensity histogram features, 25 grey-level cooccurrence matrix (GLCM) features, 16 grey level run length matrix (GLRLM) features, 16 grey level size zone matrix (GLSZM) features, 5 neighborhood grey tone difference matrix (NGTDM) features and 15 shape features.

Features were extracted in 3D using a fixed bin number of 64 and features from GLCM and GLRLM were computed from a single matrix after merging all 3D directional matrices. No harmonization in the features' space was considered in this study.

Radiomics features selection

Feature selection was performed on the training set using the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996) regression model with a survival time above or below the median survival time as a binary response variable. A 5-fold cross validation was used for the tuning parameter lambda and features were centered and scaled by subtracting the mean and divided by the standard deviation. The selected features by the model were associated with weights reflecting their significances. Feature selection was done on features obtained from H_{none}, H_{HM} and H_{GAN} images.

A radiomics score was computed using the weights of the 10 best selected features from the LASSO using the following formula:

$$\text{Radiomic score} = \sum_{i \in \text{selected features}} \omega_i . f_i ,$$

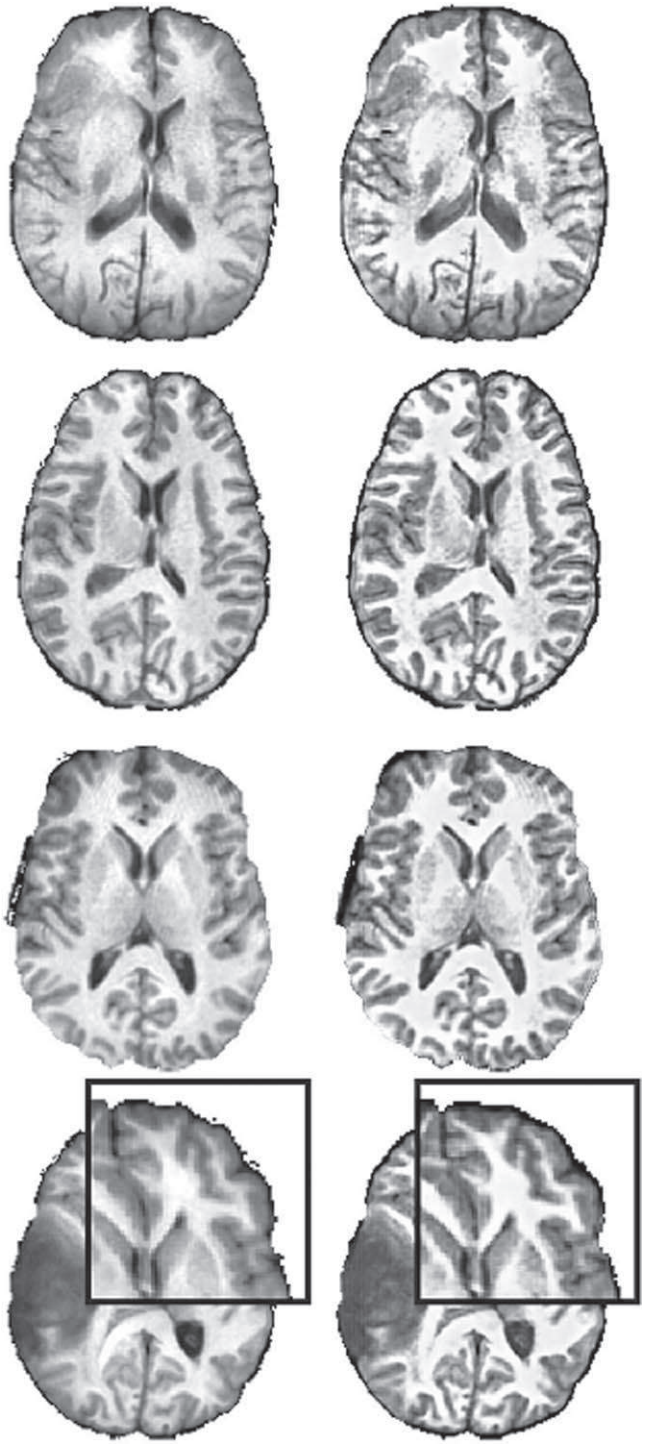

Figure 2. Example of synthesis results from H_{HM} (on the left) and H_{GAN} (on the right).

where ω_i, is the weight of the selected feature i (f_i). Weights were normalized to assign a weight of 1 to the feature with the highest contribution. The percentage contribution of a feature/feature's category (C_f) in a model was assessed by the following formula $(100 \times C_f)/W$, where W is the weight of the model (i.e. the sum of all feature weights after normalization in the model).

Radiomics OS prediction model and statistical analysis

In order to compare the performance H_{none}, H_{HM} and H_{GAN} images to build prediction models we firstly analyzed the percentage of contribution of histogram and texture (GLCM, GLRLM, GLSZM and NGTDM) features versus shape features amongst the ones selected using the H_{none}, H_{HM} and H_{GAN} images: indeed, shape descriptors are morphological descriptors solely related to the segmentation mask, therefore are not considering the intensity values of the voxels. By consequence, they cannot be impacted by the harmonization step.

Table 1. *p*-values of Wilcoxon test (adjusted with the Benjamini–Hochberg procedure) of features computed on non-standardized data versus standardized data with histogram matching and GAN. Feature with non-significant difference are bolded and features selected by LASSO are highlighted in blue.

		With H_{HM}	With H_{GAN}
HISTOGRAM	Min	<0.0001	<0.0001
	Max	<0.0001	<0.0001
	Mean	<0.0001	<0.0001
	Variance	<0.0001	<0.0001
	Standard_Deviation	<0.0001	<0.0001
	Skewness	**1**	**1**
	Kurtosis	0.0021	<0.0001
	Energy	**1**	<0.0001
	Entropy	**1**	<0.0001
	AUC	<0.0001	<0.0001
GLCM	Max	**1**	<0.0001
	Average	<0.0001	<0.0001
	Variance	<0.0001	<0.0001
	Entropy	<0.0001	<0.0001
	DAVE	<0.0001	<0.0001
	DVAR	<0.0001	<0.0001
	DENT	<0.0001	<0.0001
	SAVE	<0.0001	<0.0001
	SVAR	<0.0001	<0.0001
	SENT	<0.0001	<0.0001
	ASM	<0.0001	<0.0001
	Contrast	<0.0001	<0.0001
	Dissimilarity	<0.0001	<0.0001
	Inv diff	0.0004	<0.0001
	Inv diff norm	<0.0001	<0.0001
	IDM	0.0009	<0.0001
	IDM norm	<0.0001	<0.0001
	Inv var	<0.0001	<0.0001
	Correlation	**1**	**1**
	Autocorrelation	<0.0001	<0.0001
	Tendency	<0.0001	<0.0001
	Shade	**1**	**1**
	Prominence	<0.0001	<0.0001
	IC1	**1**	**0.4654**
	IC2	**1**	**1**

		With H_{HM}	With H_{GAN}
NGTDM	Coarseness	**1**	**1**
	Contrast	<0.0001	<0.0001
	Busyness	**1**	**1**
	Complexity	<0.0001	<0.0001
	Strength	**1**	**1**
GLRLM	SRE	0.0044	<0.0001
	LRE	**0.3337**	0.0002
	GLNU	**1**	**0.1860**
	RLNU	**1**	**1**
	RP	**0.0605**	<0.0001
	LGRE	**0.1210**	**1**
	HGRE	<0.0001	<0.0001
	LGSRE	**0.3355**	**1**
	HGSRE	<0.0001	<0.0001
	LGHRE	0.0090	**1**
	HGLRE	<0.0001	<0.0001
	GLNU norm	<0.0001	<0.0001
	RLNU norm	0.0033	<0.0001
	GLVAR	<0.0001	<0.0001
	RLVAR	**1**	0.0042
	Entropy	<0.0001	<0.0001
GLSZM	SZSE	<0.0001	<0.0001
	LZSE	**0.6077**	0.0034
	LGLZE	0.0103	**1**
	HGLZE	<0.0001	<0.0001
	SZLGE	**0.3340**	**1**
	SZHGE	<0.0001	<0.0001
	LZLGE	0.0009	0.0490
	LZHGE	**1**	**1**
	GLNU	**1**	**1**
	ZSNU	0.0022	<0.0001
	ZSP	0.0023	<0.0001
	GLNU norm	<0.0001	<0.0001
	ZSNU norm	<0.0001	<0.0001
	GLVAR	<0.0001	<0.0001
	ZSVAR	**1**	0.0223
	Entropy	**1**	**1**

Therefore, methods showing the highest percentages from the histogram and texture features could reflect the ability of harmonization techniques to improve prediction power of predictive models.

The correlation of the selected histogram and texture features with tumor core volume and selected shape features was investigated using the Pearson's correlation coefficient. This test was performed to assess if histogram and texture features selected were strongly correlated with the selected shape features and therefore used as a substitute in the model, which could potentially influence the impact of image harmonization. Also, in order to quantify the degree of influence of image harmonization on different radiomics features (histogram and texture based), a Wilcoxon paired test was carried out (*p*-values were adjusted with the Benjamini–Hochberg procedure to balance out the false discovery rate). This test was performed to verify if selected features on H_{HM} and H_{GAN} images were impacted by the harmonization.

Models were built from H_{none}, H_{HM} and H_{GAN} images using the features selection by the LASSO with (1) only the features impacted by the harmonization (this set of features also exclude shape and shape-correlated features that are not impacted by harmonization) in order to show how harmonization can improve the models and (2) all the considered radiomics features allowing to investigate if there are stable features that can substitute features impacted by the harmonization, allowing an assessment of the global performance of the models. Figure 1 shows the methodology we followed.

The ability of radiomics score to dichotomize patients into two risk groups was evaluated using the Youden J index (Youden 1950) from ROC curves and Kaplan–Meier analysis. The independent testing set was used to validate the results obtained from the training set using the thresholds found on the training set. All statistical analyses were performed using R software (v3.6.3) with caret, survival and pROC packages. *P*-values <0.05 were considered significant.

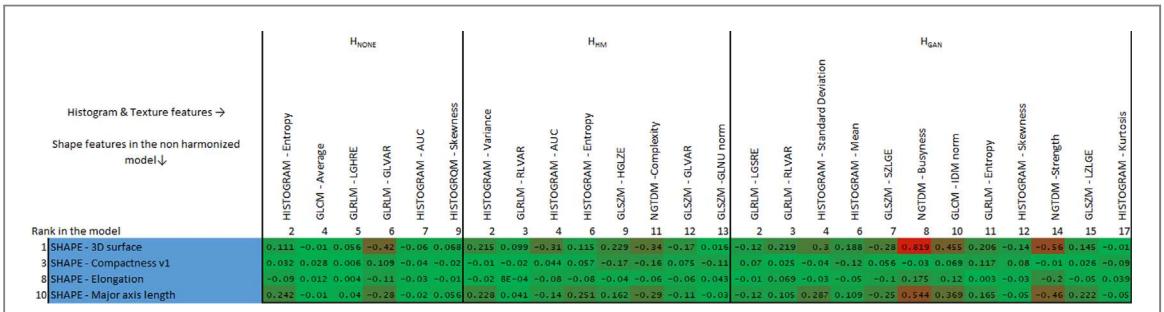

Figure 3. Correlation between shape features retrained by the H_{None} model and histogram and textural features retained on the models with the three harmonization methods.

Table 2. List of selected features by LASSO on training set with the 3 harmonization methods. Blue lines correspond of shape features (that are not impacted by the harmonization). Features weights were represented with their absolute values and normalized to assign a weight of 1 for the top ranked feature in each harmonization.

	H_{None}			H_{HM}			H_{GAN}		
Features	Ranking	Weight	Correlation with volume	Ranking	Weight	Correlation with volume	Ranking	Weight	Correlation with volume
SHAPE - 3D surface	1	1	0.95	1	1	0.95	1	1	0.95
GLRLM - LGSRE	-	-	-	-	-	-	2	0.90	-0.09
GLRLM - RLVAR	-	-	-	3	0.73	0.13	3	0.86	0.26
HISTOGRAM – Standard deviation	-	-	-	-	-	-	4	0.86	0.29
SHAPE – Least axis length	-	-	-	8	0.14	0.92	5	0.78	0.92
HISTOGRAM - Mean	-	-	-	-	-	-	6	0.74	0.17
GLSZM - SZLGE	-	-	-	-	-	-	7	0.63	-0.25
NGTDM - Busyness	-	-	-	-	-	-	8	0.43	0.84
SHAPE - Elongation	8	0.10	0.20	7	0.17	0.20	9	0.40	0.20
GLCM – IDM norm	-	-	-	-	-	-	10	0.36	0.45
GLRLM - Entropy	-	-	-	-	-	-	11	0.36	0.23
HISTOGRAM - Skewness	9	0.04	0.07	-	-	-	12	0.30	-0.13
SHAPE – Compactness v1	3	0.35	-0.004	-	-	-	13	0.17	-0.004
NGTDM - Strength	-	-	-	-	-	-	14	0.15	-0.49
GLSZM - LZLGE	-	-	-	-	-	-	15	0.14	0.15
SHAPE - Sphericity	-	-	-	5	0.34	0.002	16	0.05	0.002
HISTOGRAM - Kurtosis	-	-	-	-	-	-	17	0.02	-0.01
HISTOGRAM – Entropy	2	0.47	0.12	6	0.22	0.13	-	-	-
GLCM Average	4	0.31	0.006	-	-	-	-	-	-
GLRLM - LGHRE	5	0.20	0.08	-	-	-	-	-	-
GLRLM - GLVAR	6	0.18	-0.39	-	-	-	-	-	-
HISTOGRAM - AUC	7	0.11	-0.09	4	0.70	-0.32	-	-	-
SHAPE – Major axis length	10	0.03	0.61	10	0.11	0.61	-	-	-
HISTOGRAM - Variance	-	-	-	2	0.90	0.21	-	-	-
GLSZM - HGLZE	-	-	-	9	0.13	0.20	-	-	-
NGTDM - Complexity	-	-	-	11	0.05	-0.38	-	-	-
GLSZM - GLVAR	-	-	-	12	0.004	-0.14	-	-	-
GLSZM – GLNU norm	-	-	-	13	0.003	-0.006	-	-	-

Results

Patients

Median patient age at diagnosis was 61.5 years (range 19–87). These patients had a 1,2,3 and 5 years survival probability of 50%, 15%, 8% and 0%, respectively. Median survival was of 369 days (range 5–1767 days).

Harmonization results

Figure 2 presents synthesis results for four different images of the dataset. Visual results were more satisfying using H_{GAN} than H_{HM}, with sharp boundaries and high contrast between gray and white matter reproducing the properties of the target template image. No noticeable structure loss could be observed, suggesting the efficiency of the two-step training procedure in preserving image structures.

Radiomic features and selection

Among the 73 texture and histogram features investigated 49 were statistically different (adjusted *p*-value < 0.05) on H_{HM} images. This number increased to 55 on H_{GAN} images (table 1). In total, 57 features were found impacted on H_{HM} or H_{GAN} images (table 1).

From the 57 features found to be impacted by harmonization, the LASSO algorithm selected 14, 12 and 13 features with H_{none}, H_{HM} and H_{GAN} images, respectively. The number of selected features was of 10, 13 and 17 with H_{none}, H_{HM} and H_{GAN} images, respectively, when the LASSO regression model was applied on all the radiomic features (see table 2). The shape feature '3D surface', describing the area at the surface of the tumor volume, was selected in the three models and was associated with the highest weight. The contribution of

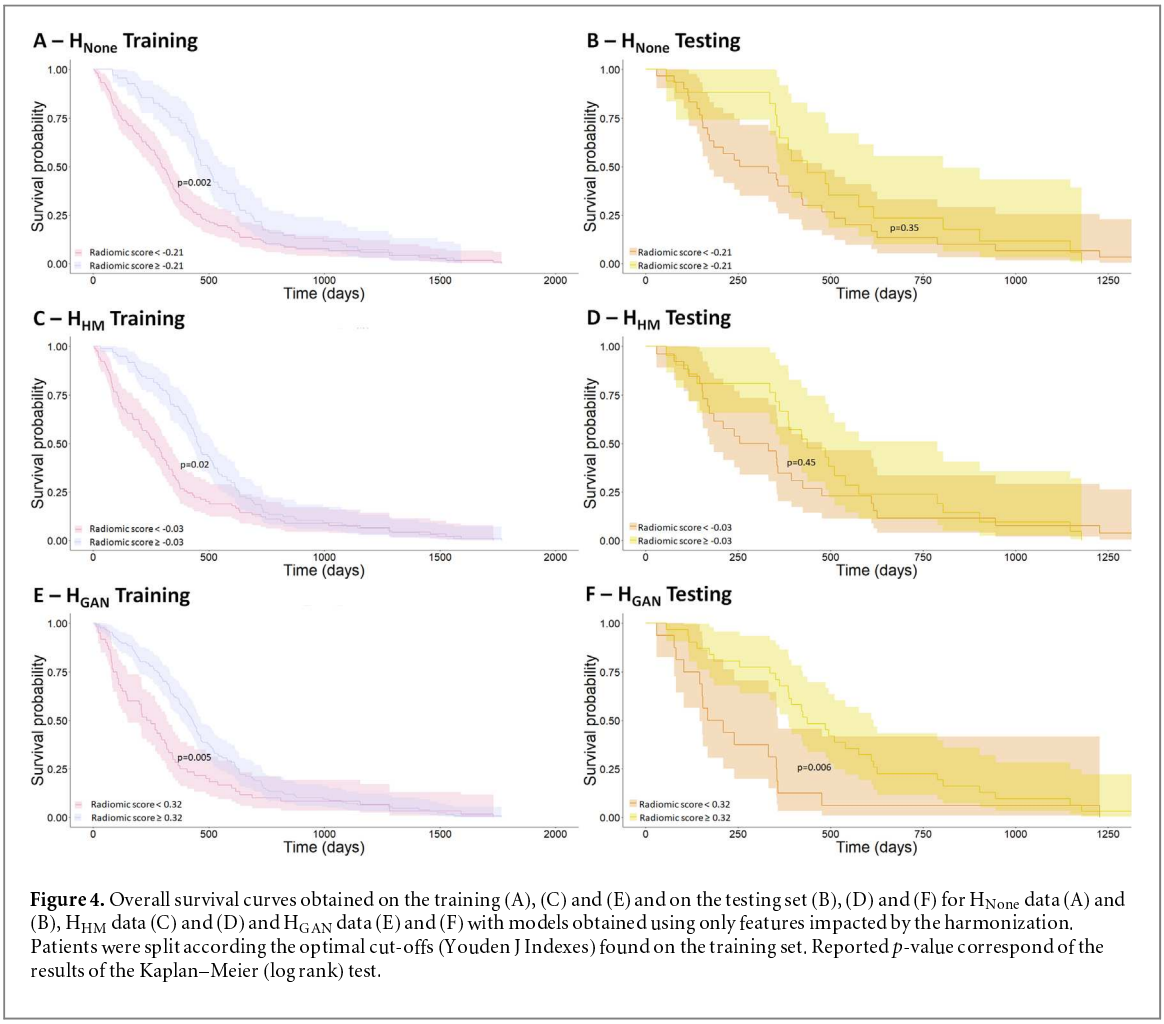

Figure 4. Overall survival curves obtained on the training (A), (C) and (E) and on the testing set (B), (D) and (F) for H_{None} data (A) and (B), H_{HM} data (C) and (D) and H_{GAN} data (E) and (F) with models obtained using only features impacted by the harmonization. Patients were split according the optimal cut-offs (Youden J Indexes) found on the training set. Reported *p*-value correspond of the results of the Kaplan–Meier (log rank) test.

Table 3. AUCs and specificity/sensitivity (Youden J Index) found on the training for the validation set with the three harmonization methods.

	Method	AUC (95% CI)	Threshold	Specificity	Sensitivity
Features impacted by the harmonization only	No harmonization	0.70 (0.54–0.85)	−0.21	0.48	0.75
	Histogram matching harmonization	0.67 (0.51–0.83)	−0.03	0.61	0.71
	GAN harmonization	0.80 (0.68–0.93)	0.32	0.91	0.58
All radiomic features	No harmonization	0.75 (0.60–0.90)	0.12	0.83	0.58
	Histogram matching harmonization	0.74 (0.59–0.89)	0.02	0.78	0.67
	GAN harmonization	0.78 (0.65–0.92)	0.04	0.65	0.67

histogram and texture features versus shape features was of 47% versus 53% for H_{none}, 62% versus 38% for H_{HM} and 71% versus 29% for H_{GAN}. The contribution of features impacted by the harmonization versus others was of 22% versus 78% for H_{none}, 39% versus 61% for H_{HM} and 41% versus 59% for H_{GAN}.

All the features from the histogram and texture matrices selected showed a weak correlation with the tumor core volume ($\rho < 0.5$), at the exception of the busyness from the NGTDM (correlation $\rho = 0.84$ (see table 2)).

Figure 3 shows correlation between the shape features used in the H_{none} model and histogram and texture features selected in the 3 different models. Again, the only high correlation was found between the 3D surface and the busyness from the NGTDM in H_{GAN} model ($\rho = 0.82$). All the other features showed non-statistically significant differences with a $|\rho| < 0.56$.

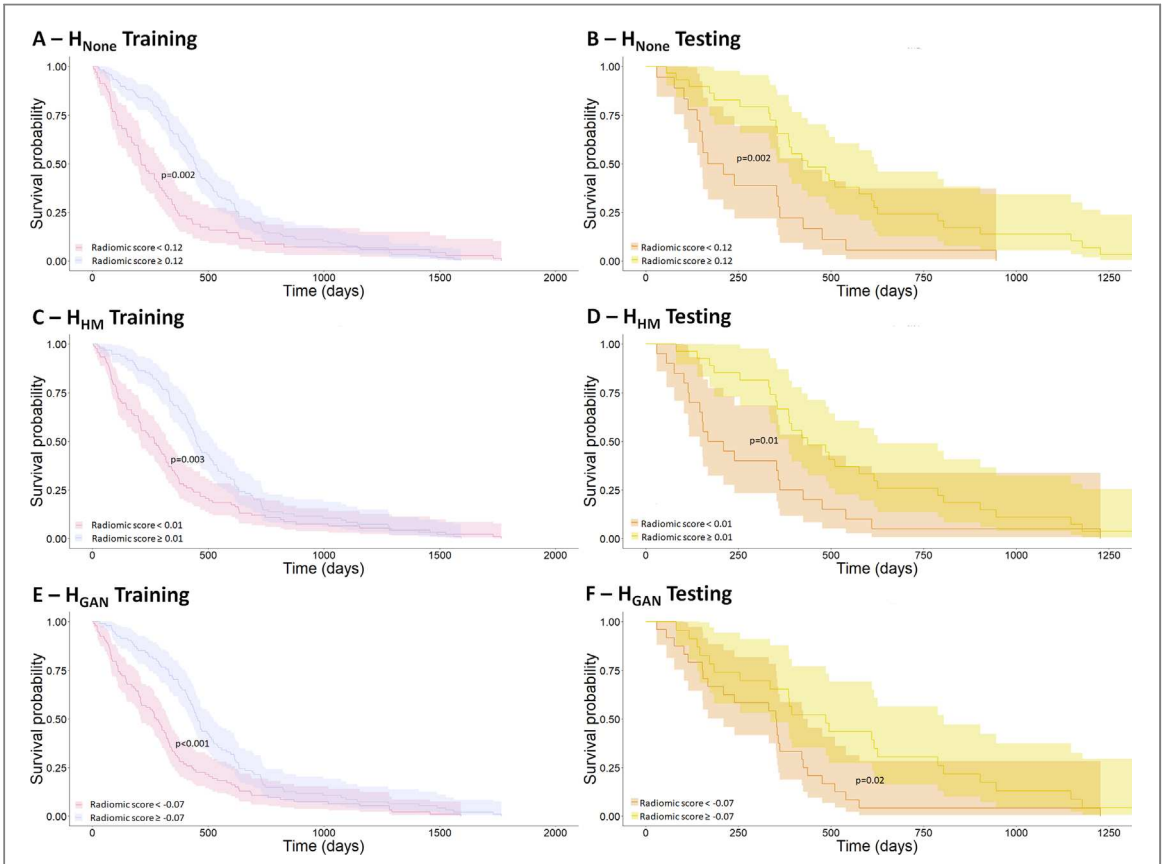

Figure 5. Overall survival curves obtained on the training (A), (C) and (E) and on the testing set (B), (D) and (F) for H_{None} data (A) and (B), H_{HM} data (C) and (D) and H_{GAN} data (E) and (F) with models obtained using the whole set of radiomic features. Patients were split according the optimal cut-offs (Youden J Indexes) found on the training set. Reported *p*-value correspond of the results of the Kaplan–Meier (log rank) test.

Prediction models

ROC curves show AUCs of 0.70 (0.54–0.85 95% CI) for H_{none} model, 0.67 (0.51–0.83 95% CI) for H_{HM} and 0.80 (0.68–0.93 95% CI) for H_{GAN} model on the testing set on the models build from features impacted by the harmonization only. Models build from the whole set of radiomic features exhibit AUCs of 0.75 (0.60–0.90 95% CI), 0.74 (0.59–0.89 95% CI) and 0.78 (0.65–0.92 95% CI) with H_{none}, H_{HM} and H_{GAN}, respectively (table 3).

On the models build from features impacted by the harmonization only, the thresholds on radiomics score which maximize the Youden J index (Youden 1950) were of −0.21, −0.03 and 0.32 for H_{none}, H_{HM} and H_{GAN} models, respectively. All three feature sets allowed to divide the patients of the training set into two groups with significantly different survival: median survival of 298 versus 495 days ($p = 0.002$) for H_{none}, median survival of 268 versus 453 days ($p = 0.002$) for H_{HM} and a median survival of 241 versus 430 days ($p = 0.006$) for H_{GAN} (figures 4(A), (C) and (E)). Only the H_{GAN} model was found able to validate the results the validation dataset with two groups of significantly different median survivals (189 versus 437 days, $p = 0.006$) (figure 4(F)). A median survival of 293 versus 437 days were found for H_{none} ($p = 0.35$) and H_{HM} ($p = 0.45$) (figures 4(B) and (D)).

On the models build from the whole set of radiomic features, the thresholds on radiomics score which maximize the Youden J index (Youden 1950) were of 0.12, 0.01 and −0.07 for H_{none}, H_{HM} and H_{GAN} models, respectively. All three datasets allowed to divide the patients of the training set into two groups with significantly different survival: median survival of 213 versus 445 days ($p = 0.002$) for H_{none}, median survival of 268 versus 448 days ($p = 0.003$) for H_{HM} and a median survival of 277 versus 454 days ($p = 0.0002$) for H_{GAN} (figures 5(A), (C) and (E)). These thresholds obtained on the training dataset allowed to dichotomize the patients on the validation dataset into two of groups with significantly different survival: median survival of 189 versus 437 days ($p = 0.003$) for H_{none}, median survival of 189 versus 437 days ($p = 0.01$) for H_{HM} and a median survival of 352 versus 486 days ($p = 0.02$) for H_{GAN} (figures 5(B), (D) and (F)).

Discussion

Although the potential usefulness of radiomics in oncology has been demonstrated in numerous studies (Kumar *et al* 2012, Lambin *et al* 2017, Hatt *et al* 2019, Rogers *et al* 2020), models are often not easy to validate in patient

cohorts outside the center where the model was initially created, if the images used differ too much from the ones used to build the model. One of the main reasons for this concerns the variability in image acquisition and reconstruction protocols which can be minimized through harmonization initiatives (Boellaard *et al* 2015, Ellingson *et al* 2015). However, although capable of reducing variability, the use of such protocol harmonization does not allow to eliminate all effects and can mostly be implemented in well controlled prospective clinical trials, which do not greatly facilitate the uptake of radiomics in routine clinical practice. In addition, manufacturers use specific image reconstruction algorithms and associated parameters and data corrections which are practically impossible to standardize. Within this context the potential of harmonization may play a key role. This can be performed in the radiomics space once the parameters have been extracted or directly in the image space. This latter solution presents multiple potential advantages for further automatization of the model building process (Visvikis *et al* 2019) but also generalization of radiomics based models to single patient datasets in a retrospective fashion. In this study we aimed at assessing if image harmonization with two different approaches (histogram matching and deep learning) can help in building radiomics models to separate patients into groups of different survival.

We were able to validate the models built on the parameters extracted using the three different image datasets (no harmonization for H_{none}, HM and GAN based for H_{HM} and H_{GAN} respectively). All these models were found to exhibit similar performance. However, the H_{HM} and H_{GAN} models included histogram and texture features that were found impacted by the harmonization process. This result clearly suggests that there are features used for the OS prediction model that are influenced by the center effect. In addition, from the models built using only features impacted by the harmonization we were only able to validate the H_{GAN} model. This result clearly suggests that advanced image harmonization has the ability to smooth the variability between images and recover the predictive power of some radiomic features. In this study, we have used the manual segmentations done by experts that are included in the BRATS dataset. Using delineation performed after image harmonization could also help to improve the prediction power of the models.

If globally, the harmonization did not significantly improve the predictive power of the OS models built using the whole set of radiomics features, these models have the advantage of selecting less tumor shape related features, which are less robust to tumor segmentation compared to histogram and texture features (Parmar *et al* 2014, Tixier *et al* 2019). Consequently, one would expect overall performance improvements for models built on harmonized data when tumors are smaller and less contrasted.

Comparing the performance of the image harmonization (H_{HM} and H_{GAN}), with feature harmonization techniques, such as Combat and its variants (Orlhac *et al* 2018, Da-ano *et al* 2020b), or investigating the combination of harmonization in both feature and image space would have been of interest but was not possible in this study given that center specific information is not available in the dataset used. In addition, the number of patients per center is too small to allow an appropriate usage of feature space harmonization techniques. In addition, even if such information was available, this dataset was built from images coming from 19 different institutions and consequently, number of patients per center would have been too small to appropriately use harmonization techniques in the feature space. Another consequence of not having the image center information available is not knowing the repartition of the image centers in the training and testing sets. However, the large number of centers guarantee the presence of heterogeneous data in both sets.

Results obtained on OS prediction models from different harmonization techniques are showing that the image harmonization in a multi-institutional cohort allows to recover the predictive value of some radiomics features that was lost due to differences in the image properties across the centers. In addition, harmonization using deep learning techniques seems to outperform the histogram matching technique, which reinforces the interest in pursuing the development of new advanced harmonization techniques for the optimized use of radiomics models in a multi-center trial setting. Given that harmonization techniques in the feature space need to assume that the distribution of each feature is the same across all the cohorts, we believe that the image harmonization techniques have the potential to outperform feature harmonization, and this will be investigated in a future study using appropriate multicenter datasets.

Q5

Acknowledgments

This study was partly funded by ANR-19-PERM-0007 POPEYE.

Conflicts of interest

The authors declare no competing interests in relation to this study.

Author contributions statement

FT, VJ and DV designed the study. CH and VJ performed the image harmonization. FT conducted the image and statistical analyses. FT, VJ and DV drafted the manuscript. FT, VJ, CH, OG, TC, DV reviewed the manuscript.

References

Bakas S *et al* 2017 Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features *Sci. Data* **4**

Bakas S *et al* 2018 Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge arXiv:1811.02629 [cited 16 May 2019]

Boellaard R *et al* 2015 FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0 *Eur. J. Nucl. Med. Mol. Imaging* **42** 328–54

Da-ano R, Visvikis D and Hatt M 2020a Harmonization strategies for multicenter radiomics investigations *Phys. Med. Biol.* (https://doi.org/10.1088/1361-6560/aba798)

Da-ano R *et al* 2020b Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies *Sci. Rep.* **10** 10248

Desseroit M-C *et al* 2016 Development of a nomogram combining clinical staging with ^{18}F-FDG PET/CT image features in non-small-cell lung cancer stage I–III *Eur. J. Nucl. Med. Mol. Imaging* **43** 1477–85

Ellingson B M *et al* 2015 Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials *Neuro-Oncol.* **17** 1188–98

Gillies R J, Kinahan P E and Hricak H 2015 Radiomics: images are more than pictures, they are data *Radiology* **278** 563–77

Gonzalez R C and Woods R E 2006 *Digital Image Processing* 3rd edn (USA: Prentice-Hall) p 976

Goodfellow I *et al* 2014 Generative adversarial nets *Adv. Neural Inf. Process. Syst.* **27** 2672–80

Hatt M, Rest C C L, Tixier F, Badic B, Schick U and Visvikis D 2019 Radiomics: data are also images *J. Nucl. Med.* **60** 38S–4S

Hatt M, Tixier F, Visvikis D and Rest C C L 2017 Radiomics in PET/CT: more than meets the eye? *J. Nucl. Med.* **58** 365–6

Hognon C, Tixier F, Gallinato O, Colin T, Visvikis D and Jaouen V 2019 Standardization of multicentric image datasets with generative adversarial networks *IEEE Nuclear Science Symp. and Medical Imaging Conf. 2019 (Manchester, United Kingdom)* [cited 2 December 2020]. https://hal.archives-ouvertes.fr/hal-02447807

Kumar V *et al* 2012 Radiomics: the process and the challenges *Magn. Reson. Imaging* **30** 1234–48

Lambin P *et al* 2012 Radiomics: EXTRACTING more information from medical images using advanced feature analysis *Eur. J. 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

Leger S *et al* 2017 A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling *Sci Rep.* **7** 13206

Lucia F *et al* 2017 Prediction of outcome using pretreatment ^{18}F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy *Eur. J. Nucl. Med. Mol. Imaging* 1–19

Menze B H *et al* 2015 The multimodal brain tumor image segmentation benchmark (BRATS) *IEEE Trans. Med. Imaging* **34** 1993–2024

Modanwal G, Vellal A, Buda M and Mazurowski M A 2020 MRI image harmonization using cycle-consistent generative adversarial network *Medical Imaging 2020: Computer-Aided Diagnosis* (Int. Society for Optics and Photonics) p 1131413 [cited 4 February 2021], https://spiedigitallibrary.org/conference-proceedings-of-spie/11314/1131413/MRI-image-harmonization-using-cycle-consistent-generative-adversarial-network/10.1117/12.2551301.short

El Naqa I *et al* 2009 Exploring feature-based approaches in PET images for predicting cancer treatment outcomes *Pattern Recognit.* **42** 1162–71

Orlhac F *et al* 2018 A post-reconstruction harmonization method for multicenter radiomic studies in PET *J. Nucl. Med.* (https://doi.org/10.2967/jnumed.117.199935)

Parmar C *et al* 2014 Robust radiomics feature quantification using semiautomatic volumetric segmentation *PLoS One* **9**

Parmar C, Grossmann P, Bussink J, Lambin P and Aerts H J W L 2015 Machine learning methods for quantitative radiomic biomarkers *Sci. Rep.* **5** 13087

Rogers W *et al* 2020 Radiomics: from qualitative to quantitative imaging *Br. J. Radiol.* **93** 20190948

Tibshirani R 1996 Regression shrinkage and selection via the Lasso *J. R. Stat. Soc.* B **58** 267–88

Tixier F, Um H, Young R J and Veeraraghavan H 2019 Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features *Med. Phys.* **46** 3582–91

Um H, Tixier F, Bermudez D, Deasy J O, Young R J and 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

Visvikis D, Cheze Le Rest C, Jaouen V and Hatt M 2019 Artificial intelligence, machine (deep) learning and radio(geno)mics: definitions and nuclear medicine imaging applications *Eur. J. Nucl. Med. Mol. Imaging* **46** 2630–7

Youden W J 1950 Index for rating diagnostic tests *Cancer* **3** 32–5

Zhong J *et al* 2020 Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development *Biomed. Eng. Online* **19** 4

Zwanenburg A, Leger S, Vallières M and Löck S 2016 Initiative for the IBS. Image biomarker standardisation initiative arXiv:1612.07003 [cited 20 November 2018]