RESEARCH ARTICLE

A transfer learning approach to facilitate ComBat-based harmonization of multicentre radiomic features in new datasets

Ronrick Da-ano[1]*, **François Lucia**[1,2], **Ingrid Masson**[1,3], **Ronan Abgral**[4], **Joanne Alfieri**[5], **Caroline Rousseau**[6], **Augustin Mervoyer**[3], **Caroline Reinhold**[7,8], **Olivier Pradier**[1,2], **Ulrike Schick**[1,2], **Dimitris Visvikis**[1]☯, **Mathieu Hatt**[1]☯

1 INSERM, UMR 1101, LaTIM, University of Brest, Brest, France, **2** Radiation Oncology Department, University Hospital, Brest, France, **3** Department of Radiation Oncology, Institut de cancérologie de l'Ouest René-Gauducheau, Saint-Herblain, France, **4** Department of Nuclear Medicine, University of Brest, Brest, France, **5** Department of Radiation Oncology, McGill University Health Centre, Montreal, Quebec, **6** Department of Nuclear Medicine, Institut de cancérologie de l'Ouest René-Gauducheau, Saint-Herblain, France, **7** Department of Radiology, McGill University Health Centre, Montreal, Canada, **8** Augmented Intelligence & Precision Health Laboratory of the Research Institute of McGill University Health Centre, Montreal, Canada

☯ These authors contributed equally to this work.
* ronrickarnaiz@gmail.com

Citation: Da-ano R, Lucia F, Masson I, Abgral R, Alfieri J, Rousseau C, et al. (2021) A transfer learning approach to facilitate ComBat-based harmonization of multicentre radiomic features in new datasets. PLoS ONE 16(7): e0253653. https://doi.org/10.1371/journal.pone.0253653

Editor: Francesco Bianconi, Universita degli Studi di Perugia, ITALY

Received: January 21, 2021

Accepted: June 9, 2021

Published: July 1, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: This study was funded by 766276 H2020 Marie Skłodowska-Curie Actions. There was no additional external funding received for this study.

Competing interests: The authors declare no competing interests.

Abstract

Purpose

To facilitate the demonstration of the prognostic value of radiomics, multicenter radiomics studies are needed. Pooling radiomic features of such data in a statistical analysis is however challenging, as they are sensitive to the variability in scanner models, acquisition protocols and reconstruction settings, which is often unavoidable in a multicentre retrospective analysis. A statistical harmonization strategy called ComBat was utilized in radiomics studies to deal with the "center-effect". The goal of the present work was to integrate a transfer learning (TL) technique within ComBat—and recently developed alternate versions of ComBat with improved flexibility (M-ComBat) and robustness (B-ComBat)–to allow the use of a previously determined harmonization transform to the radiomic feature values of new patients from an already known center.

Material and methods

The proposed TL approach were incorporated in the four versions of ComBat (standard, B, M, and B-M ComBat). The proposed approach was evaluated using a dataset of 189 locally advanced cervical cancer patients from 3 centers, with magnetic resonance imaging (MRI) and positron emission tomography (PET) images, with the clinical endpoint of predicting local failure. The impact performance of the TL approach was evaluated by comparing the harmonization achieved using only parts of the data to the reference (harmonization achieved using all the available data). It was performed through three different machine learning pipelines.

Results

The proposed TL technique was successful in harmonizing features of new patients from a known center in all versions of ComBat, leading to predictive models reaching similar performance as the ones developed using the features harmonized with all the data available.

Conclusion

The proposed TL approach enables applying a previously determined ComBat transform to new, previously unseen data.

Introduction

The extraction of quantitative features using high-throughput computing from medical images like magnetic resonance [MR], computed tomography [CT], and positron emission tomography [PET], is known as radiomics [1–4]. It provides a large set of quantitative features to researchers, enabling investigation of potential impact in clinical-decision support systems to improve diagnostic, prognostic, and predictive accuracy [5]. These various radiomics-driven prognostic/predictive studies in various cancer types may prove useful for personalized medicine in oncological applications [6].

The increased interest in radiomics research is in part due to the transparency of radiomics-based models. Thus, many initiatives have recognized the need for greater standardization of radiomics research with the aims of achieving improved reproducibility and translation of radiomics research into clinical practice [7–9]. Despite the significant impact in clinical practice, most radiomics studies to date have been single center based and retrospective in nature, and most published models have not been externally validated [10,11]. In the interest of producing convincing results with respect to the potential clinical value of radiomics as a prognostic tool, it is vital to consider large patient cohorts that can often only be available through multicenter recruitment [12–15]. One of the most important advantages of multicenter studies is the higher statistical relevance and potential generalizability of the developed models when applied to external, previously unseen cohorts. Besides the legal, ethical, administrative and technical hurdles of collecting data from several centers, one of the most challenging aspects is the fact that medical images have different characteristics when acquired on different scanner models from various manufacturers, using different acquisition protocols and reconstruction settings, which is currently unavoidable in the current clinical practice. Radiomic features have been shown to exhibit sensitivity to such heterogeneity, which consequently hinders pooling data to perform statistical analysis and/or machine learning (ML) in order to build robust models [16–21]. We recently reviewed and discussed the existing methods to perform data integration either by harmonizing images before feature extraction, or directly in the already extracted radiomics features by statistically estimating and reducing the unwanted variation associated with center effects [20,22]. In the present work, we place ourselves in the context of features harmonization (i.e., the original images are not specifically pre-processed for harmonization).

In this context, various methods have been considered [20,22]. We have recently expanded the ComBat method to improve its flexibility and robustness [23]. One remaining limitation of ComBat lies in its ability to harmonize previously unseen data (either new patients from one of the centers included in the initial harmonization process, or a new cohort from an entirely new and unseen center) [20]. In this case, the new data has to be labeled and added to the previously considered datasets, and the entire new datasets has to be re-harmonized again, which is cumbersome and seriously hinders the future external validation of the models that have been developed on harmonized features, especially if original features are not available anymore.

In this work, our objective was to develop and evaluate a transfer learning (TL) technique implemented within ComBat (and B(M)-ComBat versions as well) that could allow applying the previously learned harmonization transform to the radiomic features values of new patients from a known center.

Material and methods

ComBat approach description

The ComBat strategy was initially drafted for genomics [24], where the so-called "batch effect" is the source of variations in measurements caused by handling of samples by different laboratories, tools and technicians. This "batch effect" is theoretically similar to variations induced in radiomic features by the scanner model, the acquisition protocol and/or the reconstruction settings, sometimes called "center effect".

ComBat is primarily based on an empirical Bayes framework to eliminates batch effects. It has shown robustness with small sample sizes, down to 5 samples per batch [22,25,26], and continues to be a widely used approach [20,27–29]. ComBat was seen as being "*best* ready *to lessen and remove batch effects while expanding precision and accuracy*" when compared to five other popular batch effect removal methods [20,25]. Within the context of radiomic features harmonization, ComBat works with the following steps:

Step 1: Standardize the data

The magnitude of radiomic features could differ across center due variability in scanner models, acquisition protocols and reconstruction settings. If not accounted for, these will create bias in the Empirical Bayes (EB) estimates of the prior distribution of center effect and reduce the amount of systematic center information that can be borrowed across features [24]. To avoid this phenomenon, we first standardize the data features-wise so that radiomic features have similar overall mean and variance. Ordinary least-squares is used to calculate features-wise mean and standard deviation estimates, $\hat{\alpha}_g$ and $\hat{\sigma}_g$, across feature g, sample j, and center i. The standardized set of features Z_{ijg} from the original set of features Y_{ijg} is given by

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_{ig} - X\hat{\beta}_g}{\hat{\sigma}_{ig}} \tag{1}$$

where $X\hat{\beta}_g$ represents potential non-center related covariates and coefficients in the model.

Step 2: EB center effect parameter estimates using parametric empirical priors

The standardized data is assumed to be normally distributed $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$.

Additionally, we assume the parametric forms for prior distributions on the center effect parameters to be

$$\gamma_{ig} \sim N(Y_i, \tau_i^2) \text{ and } \delta_{ig}^2 \sim Inverse\ Gamma\ (\lambda_i, \theta_i) \tag{2}$$

The hyperparameters γ_i, τ_i^2, λ_i, θ_i are estimated empirically from standardized data using the method of moments. These prior distributions (Normal, Inverse Gamma) were selected due to the conjugacy with the Normal assumption for the standardized data [24]. Based on the distributional assumptions above, the EB estimates for center effect parameters, γ_{ijg} and δ_{ig}^2 are given (respectively) by the conditional posterior means

$$\gamma_{ig}^* = \frac{n_i\bar{\tau}_i^2\hat{\gamma}_{ig} + \delta_{ig}^{2*}\bar{\gamma}_i}{n_i\bar{\tau}_i^2 + \delta_{ig}^{2*}} \text{ and } \delta_{ig}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_j}{2} + \bar{\lambda}_i - 1} \tag{3}$$

Detailed derivations for these estimates γ_{ig} and δ_{ig}^2 are given in the supplemental materials available at *Biostatistics* online.

Step 3: Adjust the data for center effects

After calculating the adjusted center effect estimators, γ_{ig}^* and δ_{ig}^{2*}, we now adjust the data. The EB center-adjusted data y_{ig}^* is given by

$$Y_{ijg}^* = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}(Z_{ijg} - \hat{\gamma}_{ig}^*) + \hat{\alpha}_g + X\hat{\beta}_g \tag{4}$$

Although ComBat is an effective method, one of its limitations is that it centers the data to the overall, grand mean of all samples, resulting in an adjusted data matrix that is shifted to an arbitrary location that no longer coincides with the location of any of the original centers. This can result in harmonized features losing their original physical meaning, including the generation of impossible values, *e.g.*, negative volumes or SUV.

Recently, we showed the interest of two modifications [20,30]: on the one hand, a first modification, called M-ComBat, allows centering the data to the location and scale of a pre-determined "reference" batch, which, in the case of radiomics, prevents losing the physical meaning of some features (e.g., SUV or volume) and can provide the ability to select as a reference a dataset for which confidence in data curation is higher [30]. Bootstrapped ComBat (B-ComBat) on the other hand, improves the predictive ability of the developed models and their robustness through the addition of a bootstrap step [20]. These improvements however did not address the limitations of ComBat regarding the application of models based on harmonized features on new patients.

M-ComBat. M-ComBat shifts samples to the mean and variance of the chosen reference batch, instead of the grand mean and pooled variance [20,30]. This is accomplished by changing the standardizing mean and variance of the estimates, $\hat{\alpha}_g$ and $\hat{\sigma}_g$, to center-wise estimates, $\hat{\alpha}_{ig}$ and $\hat{\sigma}_{ig}$. Moreover, the mean and variance estimates utilized in the final center-effect adjusted data are calculated using the feature-wise mean and variance estimates of the reference batch, $i = r$.

The M-ComBat adjusted data are then given by

$$Y^*{}_{ijg} = \frac{\hat{\sigma}_{i=r,g}}{\hat{\delta}^*{}_{ig}}(Z_{ijg} - \hat{\gamma}^*{}_{ig}) + \hat{\alpha}_{i=r,g} + X\hat{\beta}_g \tag{5}$$

Bootstrapped ComBat: B-ComBat and BM-ComBat. Our last study [50] showed the interest of a hybrid technique performing parametric bootstrap in the initial estimates obtained in ComBat (or M-ComBat), then use a Monte Carlo method to obtain the final estimates. Hence, the final B-ComBat and BM-ComBat bootstrapped adjusted data are given respectively by:

$$Y_{ijg}^{B-ComBat} = \frac{y_{ijg} - \hat{\alpha}_{gk} - X_{ij}\hat{\beta}_{gk} - \gamma^{*}_{igk}}{\delta^{*}_{igk}} + \hat{\alpha}_{gk} + X_{ij}\hat{\beta}_{gk} \tag{6}$$

$$Y_{ijg}^{BM-ComBat} = \frac{y_{ijg} - \hat{\alpha}_{i=(r,g)k} - X_{ij}\hat{\beta}_{gk} - \gamma^{*}_{igk}}{\delta^{*}_{igk}} + \hat{\alpha}_{i=(r,g)k} + X_{ij}\hat{\beta}_{gk} \tag{7}$$

Proposed method: A transfer learning (TL) approach. We first determine the hyper parameters, *i.e*., the conditional posterior center effect estimators (γ^{*}_{ig} *and* δ^{2*}_{ig}) in the initial available dataset ("learning" part). Then, these learned hyper parameters are used in the harmonization process of the new, previously unseen dataset ("transfer" part).

The proposed method follows these steps:

i. *Save the conditional posterior center effect estimators (γ^{*}_{ig} and δ^{2*}_{ig}) obtained in **Step 2** during the initial data harmonization using ComBat.*

ii. *Perform the **Step 1** using the new unseen data.*

iii. *After obtaining the results of **(i)** and **(ii)**, perform **Step 3*** to adjust the new unseen data. The new EB center-adjusted data y^{TL}_{ijg} is given by

$$Y^{TL}_{ijg} = \frac{\hat{\sigma}_{g}}{\hat{\delta}^{*}_{ig}}(Z_{ijg} - \hat{\gamma}^{*}_{ig}) + \hat{\alpha}_{g} + X\hat{\beta}_{g} \tag{8}$$

Data: Patient cohorts, imaging and clinical endpoints

In this study, we relied on a dataset of 189 patients with histologically proven locally advanced cervical cancer (LACC) retrospectively collected from three clinical centers (Brest, n = 117 and Nantes, n = 44, in France, and Montreal, n = 28, in Canada). Patients were treated with definitive curative chemoradiotherapy followed by brachytherapy from August 2010 to July 2017 (to ensure a minimum follow-up of 1 year) (see S1 Table). The radiomics analysis was applied to the available pre-treatment images: T2-weighted MRI (T2) and apparent diffusion coefficients (ADC) maps from diffusion-weighted MRI, post-injection gadolinium contrast-enhanced MRI (GADO), and Fluorodeoxyglucose (FDG)-PET images (see S2 Table). Importantly, the PET/CT settings (scanner model, reconstruction algorithms and parameters) were the same within Brest and Nantes, but not within Montreal, where 2 different scanners were used for 5 and 23 patients respectively (see S1 Table). Compared to our previous work [20] in which 50 patients from Nantes were included, 6 were removed for the present analysis because their PET images had different characteristics. The available clinical variables included age (gender is female for all patients), histopathological type, grade, lymphovascular invasion, HPV status, T-stage, N-stage and FIGO (International Federation of Gynaecology and Obstetrics). To provide a rationale to adapt treatment (*e.g*., avoid systemic treatment for patients with low risk of recurrence), prediction of local failure (LF) was chosen as the endpoint [20,31].

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The retrospective collection of images and clinical data from the three centers was approved by the following ethics committees: "The Institutional Review Board of Brest University Hospital" for data collected in Brest and Nantes, and « The McGill University Health Center Research Ethics Board" for these collected in McGill. All patients provided permission for the use of their clinical data for scientific purposes and informed consent for the anonymous publication of data via a non-opposition form. Data were anonymized before it was accessed for the present analysis.

Overview of the framework of radiomics analysis

As in our previous work [20], we worked with radiomic features extracted from the images and the available clinical factors collected during the study by Lucia, et al [15]. These are made available for reproducibility. As a reminder, we summarize the process followed by Lucia, et al in the following. More details can be found in [15]. One single expert radiation oncologists (F. Lucia) semi-automatically delineated tumour volume-of-interests (VOIs) independently in the PET and MRI images. The Fuzzy locally adaptive Bayesian (FLAB) algorithm [32] implemented in a home-made software (MIRAS software v1.0.2.0, LaTIM, INSERM, Brest) was used in PET images while the 3D Slicer™ software [33] was used in MR images. For each VOI in the PET and three MRI sequences, 79 morphological and intensity-based features, as well as 94 textural features were extracted in 3D using MIRAS software. Features were checked for consistency with the benchmark of the Image Biomarker Standardization Initiative (IBSI) [34,35]. Both “fixed bin number” (FBN, 64 bins) and “fixed bin width” (FBW, width of 0.25, 0.5, 1 and 2 standardized uptake values (SUVs) for PET and width of 10 mm^2/s for ADC map) grey-level discretization algorithms were used to compute each of the 94 textural features. Texture matrices were built according to the merging procedure (by summation of 13 matrices calculated in toward every direction before texture calculation).

Experimental analysis

Non-parametric versions of ComBat were utilized in sections A, B, C and D using as harmonization labels either the 3 clinical centers for MRI features and the 4 scanners (1 in Brest, 1 in Nantes and 2 in Montreal) for FDG PET.

The objective of our the experiment is to demonstrate the ability of the TL implementation within ComBat to successfully harmonize features of new patients, previously not included in the initial harmonization.

Stratified random sampling was used to split the data from the 3 centers into training (n = 142 with 52 LF) and testing (n = 57 with 15 LF) sets. These sets are considered “training” and “testing” for the purpose of building multiparametric models predictive of LF using three different machine learning pipelines (see the next section). Patients from the 3 centers are thus included in both sets. For the purpose of evaluating the TL ComBat harmonization, patients from the testing set are not used in the initial harmonization, and are used to evaluate the performance of the TL harmonization. In order to evaluate the performance of TL ComBat harmonization, all patients from Nantes are set aside in the initial harmonization process and constitute the “new” center to evaluate TL harmonization. The experiment is further illustrated in Fig 1.

In order to further evaluate the impact of harmonization, principal components analysis (PCA) was performed, and the four different versions of ComBat were then compared with ANOVA in terms of their statistical distributions across labels (i.e., 3 clinical centers for MRI features and 4 scanners (1 in Brest, 1 in Nantes and 2 in Montreal) for FDG PET) before and after harmonization with the four ComBat versions. In addition, a 2-sample Kolmogorov-

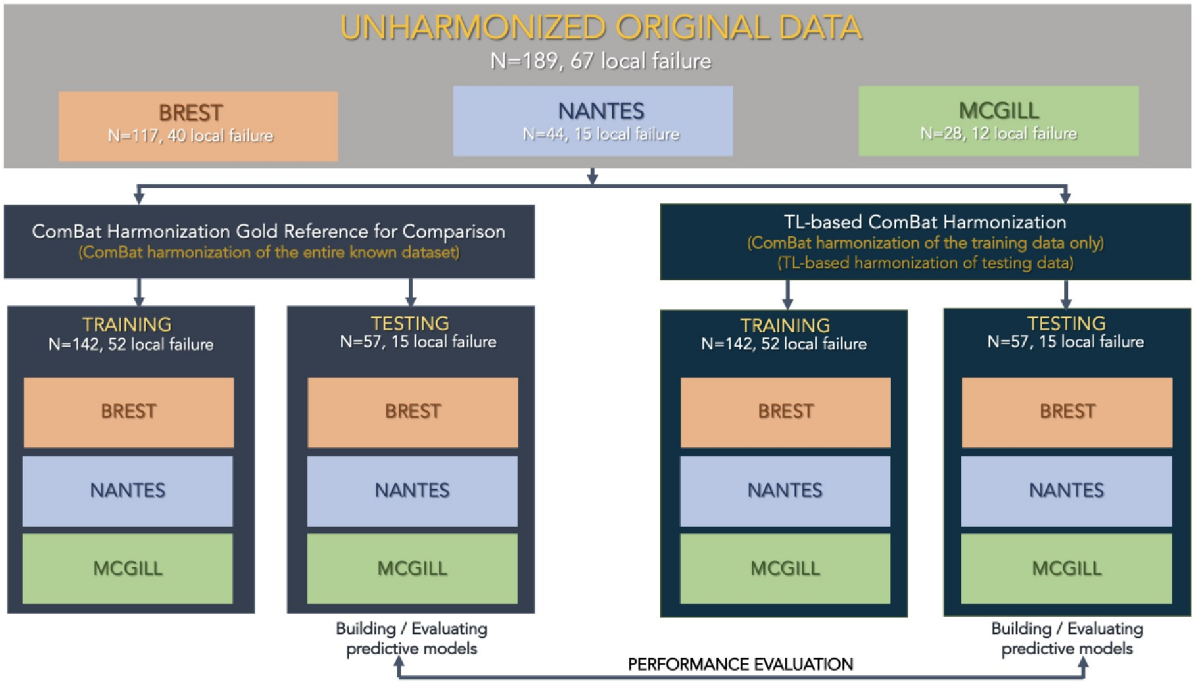
UNHARMONIZED ORIGINAL DATA
N=189, 67 local failure
BREST
N=117, 40 local failure
NANTES
N=44, 15 local failure
MCGILL
N=28, 12 local failure
ComBat Harmonization Gold Reference for Comparison
(ComBat harmonization of the entire known dataset)
TL-based ComBat Harmonization
(ComBat harmonization of the training data only)
(TL-based harmonization of testing data)
TRAINING
N=142, 52 local failure
TESTING
N=57, 15 local failure
BREST
NANTES
MCGILL
Building / Evaluating predictive models
PERFORMANCE EVALUATION

Fig 1. Workflow for the analysis in LACC datasets' experiments.

https://doi.org/10.1371/journal.pone.0253653.g001

Smirnov test was used to determine if there was a significant difference in the distribution of the features from each device variations both before and after ComBat harmonization.

The Original (i.e., original harmonized data with all the available samples, thus the "gold standard reference") was compared to the harmonized data by the proposed TL (i.e., harmonized data by the proposed transfer learning) method. Since variances of the results acquired by the two experiments could not be considered equal, we used the Welch-t test [26] to compare whether the differences between the original (i.e., gold standard) and the TL results were statistically different. Tests were run on each combination of data under the null hypotheses "method does not impact ML performances" and "both methods have same performances" and reject the Ho if $p < 0.05$.

Finally, the performance of the predictive models built relying on features harmonized with the TL approach (through all four ComBat versions) was compared to the performance of the models built using features harmonized using all the available data.

Models predicting endpoints (as a binary task) were built exactly as in our previous work: 3 different ML pipelines were utilized: i) Random Forest (RF) and ii) Support Vector Machine (SVM), both with embedded feature selection, and iii) Multivariate regression (MR) with 10-fold cross-validation after feature selection based on least absolute shrinkage and selection operator (LASSO). All of the harmonized (with the 4 ComBat versions) radiomic features were used as inputs in combination with the available clinical factors (age, gender, histology, stage, etc.). Since there were ~34% of events (LF), we used synthetic minority over-sampling technique (SMOTE) to facilitate training of the models [20,36].

For the purpose of the M-ComBat and BM-ComBat, Brest was chosen as the reference to which the two other centers (in the case of MRI) or 3 other scanners (in the case of PET) were harmonized. Fig 2 illustrates the overall workflow.

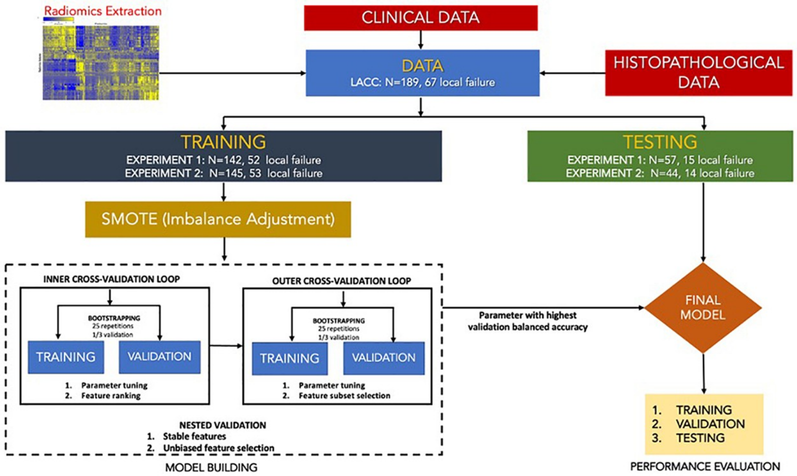
Radiomics Extraction
CLINICAL DATA
DATA
LACC: N=189, 67 local failure
HISTOPATHOLOGICAL DATA
TRAINING
EXPERIMENT 1: N=142, 52 local failure
EXPERIMENT 2: N=145, 53 local failure
TESTING
EXPERIMENT 1: N=57, 15 local failure
EXPERIMENT 2: N=44, 14 local failure
SMOTE (Imbalance Adjustment)
INNER CROSS-VALIDATION LOOP
BOOTSTRAPPING
25 repetitions
1/3 validation
TRAINING
VALIDATION
1. Parameter tuning
2. Feature ranking
OUTER CROSS-VALIDATION LOOP
BOOTSTRAPPING
25 repetitions
1/3 validation
TRAINING
VALIDATION
1. Parameter tuning
2. Feature subset selection
NESTED VALIDATION
1. Stable features
2. Unbiased feature selection
MODEL BUILDING
Parameter with highest validation balanced accuracy
FINAL MODEL
1. TRAINING
2. VALIDATION
3. TESTING
PERFORMANCE EVALUATION

Fig 2. Overall workflow for the analysis in LACC datasets.

https://doi.org/10.1371/journal.pone.0253653.g002

Machine learning methodology

Imbalance adjustment. To address the imbalance in the dataset, SMOTE algorithm was utilized to facilitate training of the models. SMOTE is a method of over-sampling the minority class in order to provide a balanced number of positive and negative cases to the learning algorithm [20,36]. The difference of SMOTE to any other techniques is that the minority class is over-sampled by creating a synthetic sample rather than over-sampling with replacement [20,36].

Multivariate regression with LASSO. Features selected from LASSO was using for training multivariate regression. Here, LASSO was used as both a regularization and variable selection methods for any statistical models [20,37]. LASSO was used to penalize the negative log of the partial likelihood in multivariate cox regression [37]. The algorithm employs a cyclical coordinate descent, which sequentially optimizes the objective function over a parameter with others kept fixed, and cycles repeatedly until convergence [20,37].

Random forest. The RF method is designed to make use of an ensemble method consisting of many decision trees [20,38]. The concept behind RF is that each decision tree is formed by choosing the sample from the original dataset with the bootstrap method and selecting the random number of all variables in every decision node. The RF strategy consists of the following steps: i.) n features are randomly selected from a total of m features, ii.) the node d the used to best split point is calculated using the n features, iii.) it checks whether the number of final nodes reaches the target number, and iv.) by repeating step i to iii for n times, a forest is then built [20,38].

Support vector machine. SVM is a supervised learning algorithm was incorporated based on a statistical learning theory [39]. It works by aiming to find the hyper-plane, which separates classes from each other, and which is the most distant from both classes. The result is a linearly separable dataset made by using a kernel function [20,39]. Also, a non-linear separation can be made, and the data can be separated in the high dimensions [39] which sometimes resulted to over-fitting in the input space. Overfitting is controlled through the principle of structural risk minimization [20,39].

Feature selection methods. The objective of feature selection is to improve the prediction performance of the predictors, understand the underlying process that generated the data and in most of the cases to provide faster and more cost-effective predictors. Given a training data set consisting of N instances, P predictor variables/features X_i (i = 1,. . ., P) and the class Y in {1, 2,. . ., C}, the objective of feature selection is to select a compact variable/feature subset without loss of predictive information about Y. Feature selection were embedded in both RF and SVM as a part of the model training process and hyper parameters optimization. It is in this manner typically specific to a given learning algorithm, *i.e*., the feature subset selection can be considered as a search in the combined space of feature subsets and hypotheses [20,40]. Regarding RF, a single decision partitions the input space into a set of disjoint regions, and assigns a response to each corresponding region [20,40]. In the event of SVM, a similar procedure was thought of in spite of the fact that, rather than the measure of variable importance as in RF, features are ranked based on the best fine cost of the models and are ranked according to the values of leave-one-out error (LOO−*i*), *i.e*., the feature *i* with the highest value of LOO−*i* is ranked first [20,41].

Final model construction and evaluation. Multiple regression with LASSO, RF and SVM models were fitted independently with the selected optimal features subset and parameters to exploit the feature selection and parameter tuning results.

For classification problem, the discrimination evaluation of the optimal solution during the training can be defined on different performance matrix. Sensitivity, Specificity and accuracy results are provided in the S1–S3 Figs. We have decided to focus on three metrics, area under the ROC curves (AUC), balanced accuracy (BAcc) and Matthew's correlation coefficient (MCC, worst value = -1; best value = +1) [20,42] to compare the results obtained without and with the different ComBat versions. ROC-AUC measure the optimal learning model underneath the ROC curve which AUC values reflects the overall ranking performance of a classifier based on thresholding settings. BAcc (calculated as the average of sensitivity and specificity) is a more appropriate metric in the presence of data imbalance than the conventional accuracy. Lastly, MCC is a contingency matric method calculating the Pearson product-moment correlation coefficient between the actual and predicted and is a good metric to measure the quality of the binary classification.

Results

Initial analysis

The COV measurements (Table 1) show that in the testing sets of the experiment, the TL data exhibit similar variability in all ComBat-harmonized versions (slightly lower/higher in ComBat and B-ComBat, slightly higher/lower in M-ComBat and BM-ComBat) as the one observed in harmonization carried out using all the data.

According to ANOVA, 97% and 98% (in *Orig* and *TL*, respectively) of untransformed radiomic features were significantly (at $p<0.01$ level) different between labels. After harmonization, all of the four ComBat versions (in both Orig and TL) completely eliminated significant label related differences across the different cohorts in both datasets, *i.e*., none of the radiomic features remained significantly (at $p<0.01$ level) different between labels.

Table 2 confirms that any ComBat versions and the untransformed data have a significant difference in data distribution. However, for both experiments 1 and 2, the four ComBat versions in the Original and TL, respectively have similar data distributions.

Scatterplots of the top two principal components of PCA (Figs 3 and 4, representing ~54% and ~53% of the information in Orig and TL, respectively) confirm visually the ability of all four ComBat versions in removing the differences in radiomic features between labels while shifting the data to different locations. In the case of TL scenario, it shows that the

Table 1. COV computed on the Orig and TL data in four ComBat versions.

Data	COV	
	Original	TL
Untransformed	2833	2833
ComBat	1313	1159
B-ComBat	1290	1082
M-ComBat	1204	1309
BM-ComBat	1189	1210

Original = original harmonized data (gold standard reference), TL = harmonized data by the proposed transfer learning method.

https://doi.org/10.1371/journal.pone.0253653.t001

Table 2. P-values of Kolmogorov-Smirnov comparing different 4 ComBat versions using Original and TL data.

Kolmogorov-Smirnov				
Untransfomed *vs*. any ComBat	Original *vs*. Transfer Learning			
	ComBat	B-ComBat	M-ComBat	BM-ComBat
<0.0001	0.753	0.921	0.977	0.992

https://doi.org/10.1371/journal.pone.0253653.t002

method is rather effective in eliminating confounding effects brought by aforementioned variabilities.

Predictive modelling using machine learning approaches

Table 3 provides results for the 3 performance evaluation metrics in the testing sets, for considering the use of the different ML algorithms in combination with the two different testing datasets, using the 4 versions of ComBat. The same results (including training sets and additional evaluation metrics) are provided in S1–S3 Figs.

The experiment involves stratified random sampling used to split the data from the 3 centers into training and testing sets. In both sets, patients from the 3 centers are thus present. For the purpose of evaluating the TL ComBat harmonization, all patients from the testing set are not used in the initial harmonization. The results indicated in Table 3 show that all machine learning approaches led to models with good predictive performance (AUC 0.84–0.95, BAcc 81–93% and MCC 0.64–0.86), in the gold standard reference. All ComBat harmonized sets of features allowed for better models than using the untransformed data. In the TL scenario, the new patients (a set containing patients from all 3 centers) could be harmonized efficiently based on the learning of the harmonization transform, as the performance of the resulting model is very similar to the one of the models obtained using all the available data: between -0.06 to +0.02 in AUC, -2 to +4% in BAcc and -0.06 to +0.06 in MCC. Consistent with the gold reference, the absolute increase in performance between the use of the original, untransformed features and the harmonized ones utilizing the TL approach changed slightly depending on the ML algorithms utilized and between the patient populations considered. Table 4 confirms a statistically significant improvement for the three ML classification methods after harmonization compared to the use of untransformed features (in both *Original* and *TL*, respectively). Moreover, Table 4 also shows the lack of significant difference between the performance of ML models in original *vs*. TL.

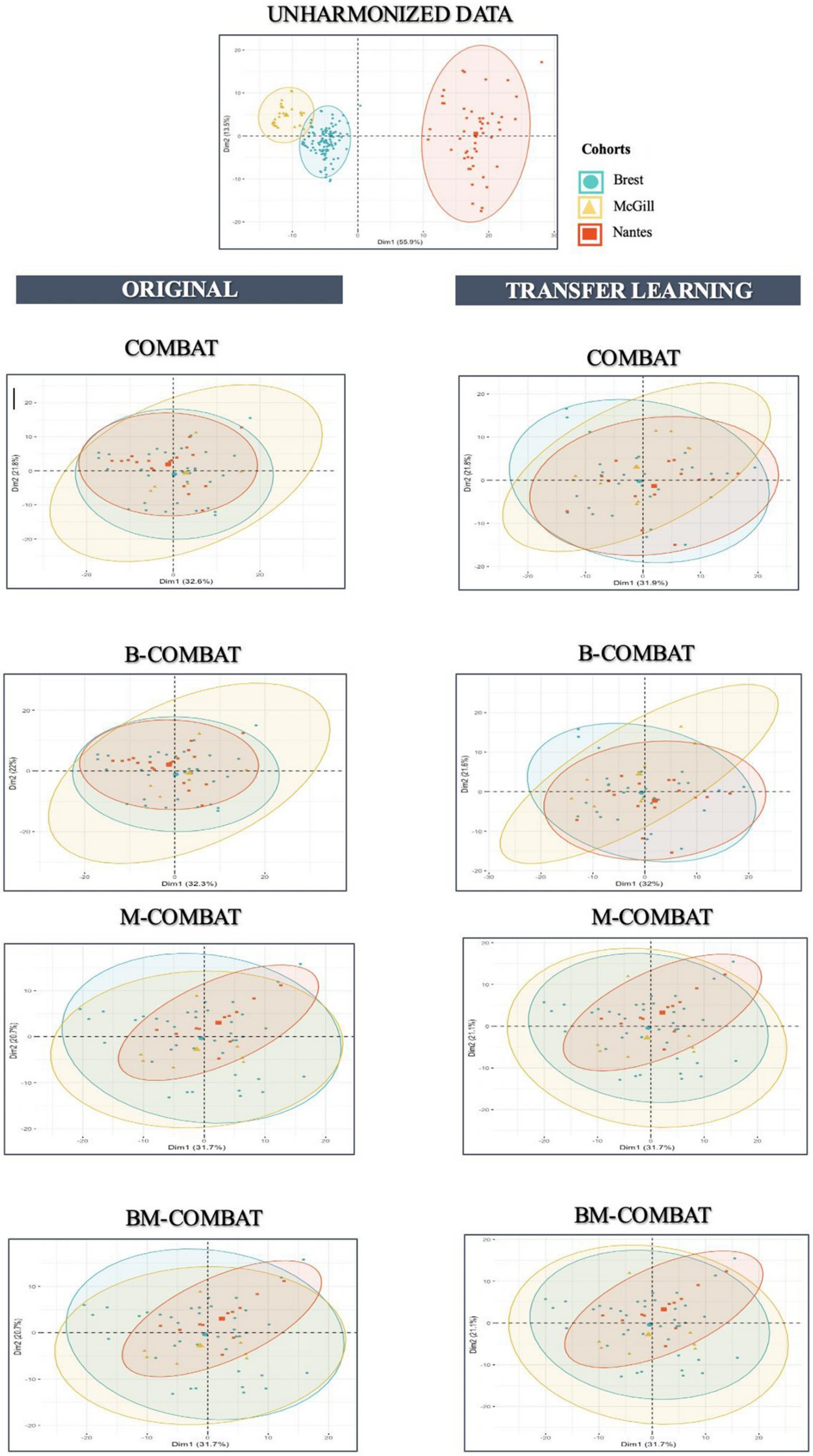

Fig 3. PCA and summary distribution in Experiment 1 in MRI. Scatter plots of top 2 principal components of the radiomic features across the three labels (centers) using data transformed with the 4 versions of ComBat (using R (3.5.1) and R Studio (1.1.456,R Studios Inc., Boston,MA) https://cran.r-project.org/).

https://doi.org/10.1371/journal.pone.0253653.g003

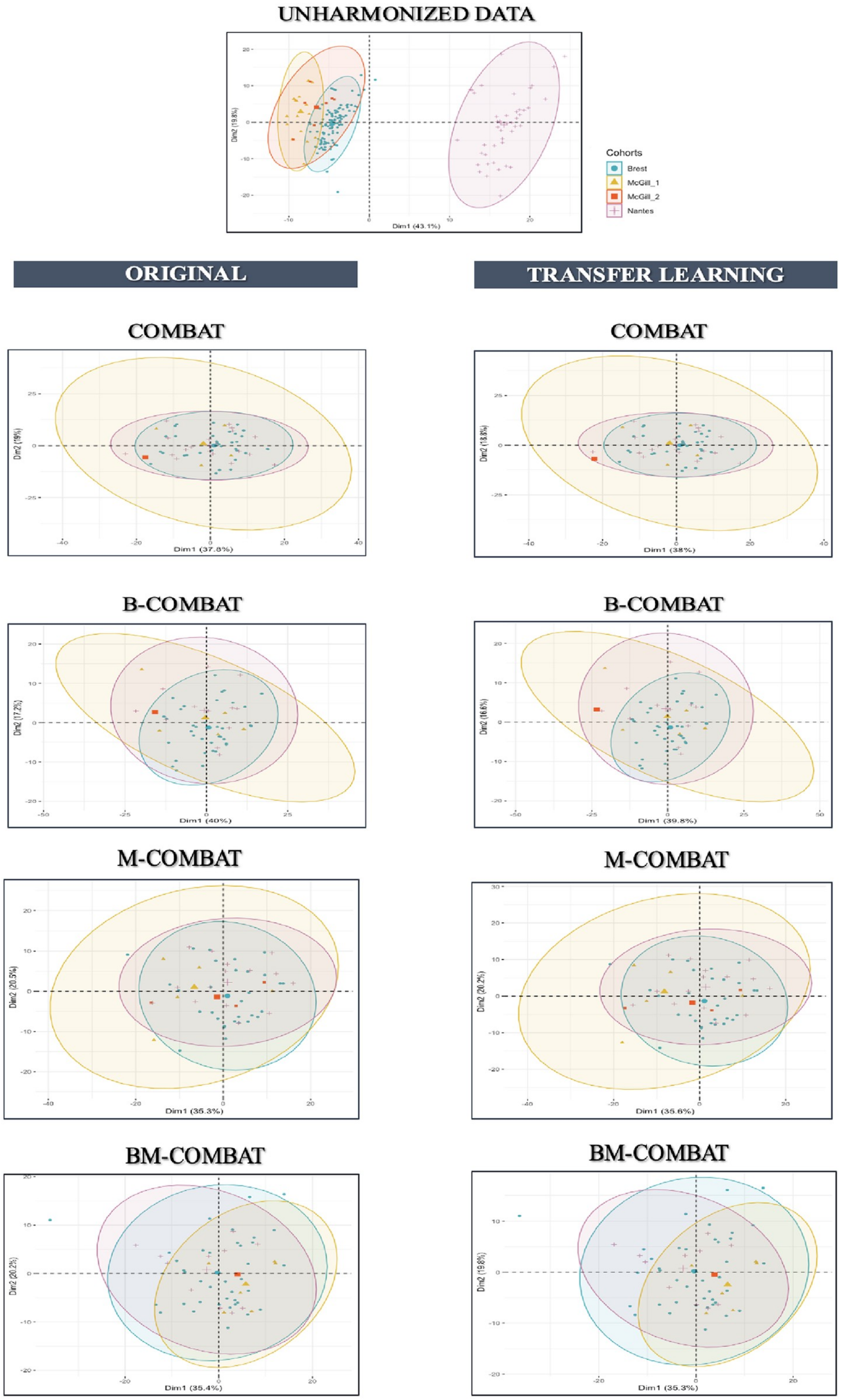

Fig 4. PCA and summary distribution in Experiment 1 in FDG PET. *Scatter plots of top 2 principal components of the radiomic features across the three labels (centers) using data transformed with the 4 versions of ComBat (using R (3.5.1) and R Studio (1.1.456,R Studios Inc., Boston,MA)* https://cran.r-project.org/).

https://doi.org/10.1371/journal.pone.0253653.g004

Table 3. Performance metrics evaluation of predictive models in testing sets using the three ML pipelines.

ML	Data	AUC [0,1]		BAcc (%)		MCC [-1,+1]	
		Original	TL	Original	TL	Original	TL
MR	Untransformed	0.80	0.80	79	79	0.48	0.48
	ComBat	0.86	0.86	84	83	0.64	0.58
	B-ComBat	0.89	0.89	87	88	0.76	0.71
	M-ComBat	0.89	0.83	83	87	0.71	0.73
	BM-ComBat	0.90	0.91	85	88	0.73	0.78
RF	Untransformed	0.84	0.84	82	82	0.67	0.67
	ComBat	0.91	0.90	87	86	0.73	0.70
	B-ComBat	0.93	0.94	91	92	0.82	0.86
	M-ComBat	0.90	0.92	88	88	0.81	0.83
	BM-ComBat	0.93	0.95	93	94	0.86	0.90
SVM	Untransformed	0.79	0.79	77	77	0.61	0.61
	ComBat	0.85	0.83	81	80	0.66	0.65
	B-ComBat	0.88	0.88	83	84	0.73	0.75
	M-ComBat	0.90	0.89	83	81	0.70	0.76
	BM-ComBat	0.91	0.93	85	86	0.73	0.79

Original = original harmonized data (gold standard reference), TL = harmonized data by the proposed transfer learning method.

https://doi.org/10.1371/journal.pone.0253653.t003

Table 4. P-values of Welch's t-test comparing ML algorithms performances on different 4 ComBat versions using original and TL data.

ML	Untransfomed vs. any ComBat	Original vs. Transfer learning			
		ComBat	B-ComBat	M-ComBat	BM-ComBat
MR	<0.0001	0.15	0.13	0.16	0.15
RF	<0.0001	0.14	0.12	0.15	0.13
SVM	<0.0001	0.13	0.12	0.15	0.13

https://doi.org/10.1371/journal.pone.0253653.t004

Discussion

Variations of scanner models, reconstruction algorithms and acquisition protocols are frequently unavoidable in multicenter studies, just as in retrospective studies with a long enlistment duration (*e.g.*, when the scanner is replaced by one more model at some point). In such a context, there is a clear need for harmonization in order to allow for efficient models to be trained and validated. There are two main approaches to address this issue: (i). harmonizing images (i.e., before extracting features) and (ii). harmonizing features (i.e., *posteriori*, after their extraction). The first method tackles the issue in the image domain and early developed approaches considered standardization of acquisition protocols and reconstruction settings, relying on guidelines already available, e.g., for PET/CT imaging [43,44]. However, it has been shown recently that although such an approach can help towards reducing multicentre effects, it may still be insufficient to fully compensate them [43,45]. Techniques based on deep learning (convolutional neural networks, CNN or generative adversarial networks, GAN and their variants) have also been considered in order to standardize or harmonize medical images [44,46–48], including with an evaluation of the impact on resulting radiomic features, in the context of lung lesions in CT images [46]. Another paper [49] showed in a proposed workflow evaluating harmonization techniques using synthetic and real data comparing ComBat and cycleGaN that both methods perform well for removing various types of noises while preserving manually added synthesis lesions, but also for removing site effects on data coming from 2 different sites while preserving biological information. These techniques are promising but do not appear mature enough yet to enable a full comparison with the harmonization in the feature's domain. Thus, as our team is currently developing such methods [50], we will definitively carry out such comparisons in future studies.

The other approach addresses the issue in the feature domain. This can be done either i) by selecting features before the statistical analysis based on their robustness, in order to eliminate features too sensitive to multicentre variability, or ii) by retaining all features together with their harmonizing their statistical properties so they can be grouped throughout the modeling step [20]. Numerous statistical methods exist to perform such normalization or batch-effect correction [20,26]. ComBat recently outperformed 6 other methods for batch effect removal using microarray datasets from brain RNA samples and two simulated datasets [25]. Although an extensive comparison of ComBat with other methods remains to be carried out explicitly in the context of radiomics, it has already been identified as a promising technique and is being increasingly and successfully used in recent radiomics studies [15,20,22,24,32,51–56]. It however has some limitations regarding its use in practice and we previously addressed two of these with the proposed BM-ComBat to allow for more flexibility in choosing a reference label and improving the estimation [44]. As expected and similarly to previous findings [44], in the experiment, all versions of ComBat were able to remove the differences amongst radiomic features and improve the predictive performance of the models, and the best results were consistently obtained with B(M)-ComBat over the standard ComBat, whether in the context of Original or TL scenarios.

The magnitude of differences between the performance of models trained and evaluated either in the original or the TL scenario is similar to the differences observed in a given scenario between different ML approaches. This absolute difference in performance amongst the ML algorithms can be attributed in part to i) the different feature selection techniques [41] and ii) the way the classifiers combine selected features. Previous studies have also highlighted the variability of resulting performance of radiomic models depending on either classifier or feature selection algorithms [20,57,58].

We proposed and evaluated a transfer learning modification to the well-known ComBat methodology for eliminating center-effects that allowed transferring the previously learned harmonization transform to the radiomic features based signatures values of new patients from a known center. Principal components analysis, analysis of variance, and statistical tests have shown the feasibility of this proposed harmonization approach, in the sense that the efficiency of the harmonization and the resulting performance of trained models in testing dataset is similar with the proposed TL approach, compared to the reference gold standard using all available data for the harmonization. These demonstrated that the proposed TL technique leads to efficient estimation with similar resulting predictive ability of models. This important point was demonstrated across 3 different ML algorithms, all performance metrics and for both experiments. More specifically, the experiment showed that the TL approach was effective in applying the previously determined harmonized transform to the radiomic features values of new patients from a known center resulting in a consistent improvement in the predictive performance of the developed models. Although the proposed TL technique provided a consistent comparable predictive performance of the developed models in different ML algorithms, we acknowledge the limitations associated with relatively small improvements in combination with a single dataset with limited heterogeneity in the imaging factors. We also performed a single split of the data and did not investigate different combinations of training/testing with the 3 available centers, due to the time-consuming building of numerous models for evaluation. Future work could consider different splits and combination of centers. The proposed method will thus require validation in larger and more diverse cohorts (more centers, more scanners and sources of variability). Our future work will thus consider the use of a small set of patients from the entirely new center as examples to learn from, in order to improve the performance of the proposed TL approach in this context. In addition, our proposed approach does not alleviate one of its inherent limitations: (i) ComBat only works properly when available and labelled data is available in order to perform the estimate and batch correction, (ii) in order to apply the developed/validated model (*i.e.*, a combination of harmonized radiomic features with an associated threshold value) to a new patient from another center not previously included, there is currently no direct method to apply the previously determined harmonization transform to the radiomic features values of this new patient in order to determine his/her prediction.

Finally, as in our previous work, we have considered working with the entire set of radiomic features irrespectively of their robustness (i.e., without first selecting features based on their resilience to changes in reconstruction or acquisition settings). Identifying radiomic features robust to changes in acquisition and reconstruction settings prior to feeding them to the machine learning pipeline is also a different approach. Such a feature selection procedure can help building more robust models, potentially without the need for harmonization, since only features insensitive (or at least, less sensitive) to multicenter variability are therefore exploited. This approach however may suffer from a potential loss of information, as features identified as unreliable are usually discarded before being evaluated and the most robust/reproducible features might not necessarily be the most clinically-relevant to the task at hand. Furthermore, the size of the radiomic features set would depend on the chosen threshold of what is considered robust enough. We will compare such an approach with ComBat harmonization in our future works.

None of the models predicting local failure selected shape features and they relied only on intensity and textural ones (whether considering the untransformed or the harmonized features), which indicates that at least in that application, the shape and size of the tumor in PET and MRI is not informative, as already observed in our initial studies in that cohort [32,59]. Shape features can be expected to be less impacted by the center effect, compared to intensity and textural features, especially since the delineation was the same for all images. However, they might still be sensitive to factors such as spatial resolution (more or less blur at the edges of the tumors will drive more or less complex shapes and surfaces) and voxel sampling (larger voxels will lead to less detailed delineations and “simpler” shapes and surfaces). Indeed, distributions of most shape features were found to be statistically different between the 3 MRI or the 4 PET batches, although the statistics was lower than for intensity and textural features.

Finally, it could also be interesting to investigate the feasibility to apply this transfer learning approach in a different implementation framework such as distributed learning [60,61].

Conclusion

The transfer learning technique implemented within ComBat allowed applying the previously determined harmonization transform to the radiomic features values of new patients from a known center with a slightly stronger decrease in performance. Our approach alleviates one of the most important limitations of ComBat for harmonization of radiomic features in a multicentre context when new, previously unseen data are to be analyzed.

Supporting information

S1 Fig. Performance metrics evaluation of predictive models using MR with LASSO.
(TIFF)

S2 Fig. Performance metrics evaluation of predictive models using RF.
(TIFF)

S3 Fig. Performance metrics evaluation of predictive models using SVM.
(TIFF)

S1 Table. Patient's characteristics.
(PDF)

S2 Table. PET/CT and MRI protocols in Brest (A), Nantes (B), and McGill (C).
(PDF)

S1 Data.
(CSV)

Author Contributions

Conceptualization: Ronrick Da-ano, Dimitris Visvikis, Mathieu Hatt.

Data curation: François Lucia, Ingrid Masson, Ulrike Schick.

Formal analysis: Ronrick Da-ano.

Investigation: Ronrick Da-ano.

Methodology: Ronrick Da-ano, Dimitris Visvikis, Mathieu Hatt.

Software: Ronrick Da-ano.

Supervision: Dimitris Visvikis, Mathieu Hatt.

Validation: Ronrick Da-ano.

Visualization: Ronrick Da-ano.

Writing – original draft: Ronrick Da-ano, Joanne Alfieri, Caroline Rousseau, Augustin Mervoyer, Caroline Reinhold, Olivier Pradier, Ulrike Schick.

Writing – review & editing: Ronrick Da-ano, Ronan Abgral, Joanne Alfieri.

References

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb; 278(2):563–77. https://doi.org/10.1148/radiol.2015151169 Epub 2015 Nov18. PMID: 26579733.
2. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information frommedical images using advanced feature analysis. Eur J Cancer. 2012 Mar; 48(4):441–6. https://doi.org/10.1016/j.ejca.2011.11.036 Epub 2012 Jan 16. PMID: 22257792.
3. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012 Nov; 30(9):1234–48. https://doi.org/10.1016/j.mri.2012.06.010 Epub 2012 Aug 13. PMID: 22898692.
4. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb; 278(2):563–77. https://doi.org/10.1148/radiol.2015151169 Epub 2015 Nov 18. PMID: 26579733.
5. Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol. 2017 Feb; 90 (1070):20160665. https://doi.org/10.1259/bjr.20160665 Epub 2016 Dec 12. PMID: 27936886.
6. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012 Mar; 48(4):441–6. https://doi.org/10.1016/j.ejca.2011.11.036 Epub 2012 Jan 16. PMID: 22257792.

7. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present. . . any future? Eur J Nucl Med Mol Imaging. 2017 Jan; 44 (1):151–165. https://doi.org/10.1007/s00259-016-3427-0 Epub 2016 Jun 6. PMID: 27271051.

8. Sollini M, Cozzi L, Antunovic L, Chiti A, Kirienko M. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. Sci Rep. 2017 Mar 23; 7(1):358. https://doi.org/10.1038/s41598-017-00426-y PMID: 28336974.

9. Nyflot M. J., Yang F., Byrd D., Bowen S. R., Sandison G. A., & Kinahan P. E. (2015). Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *Journal of medical imaging (Bellingham, Wash.)*, 2(4), 041002. https://doi.org/10.1117/1.JMI.2.4.041002 PMID: 26251842

10. Li ZC, Bai H, Sun Q, Li Q, Liu L, Zou Y, et al. Multiregional radiomics features from multiparametric MRI for prediction of MGMT methylation status in glioblastoma multiforme: A multicentre study. Eur Radiol. 2018 Sep; 28(9):3640–3650. https://doi.org/10.1007/s00330-017-5302-1 Epub 2018 Mar 21. PMID: 29564594.

11. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiother Oncol. 2018 Jun; 127 (3):370–373. https://doi.org/10.1016/j.radonc.2018.03.004 Epub 2018 Mar 26. PMID: 29598835.

12. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017 Dec; 14 (12):749–762. https://doi.org/10.1038/nrclinonc.2017.141 Epub 2017 Oct 4. PMID: 28975929.

13. Sun C, Tian X, Liu Z, et al. Radiomic analysis for pretreatment prediction of response to neoadjuvant chemotherapy in locally advanced cervical cancer: A multicentre study. *EBioMedicine*. 2019; 46:160–169. https://doi.org/10.1016/j.ebiom.2019.07.049 PMID: 31395503

14. Dissaux G, Visvikis D, Da-Ano R, Pradier O, Chajon E, Barillot I, et al. Pretreatment ^{18}F-FDG PET/CT Radiomics Predict Local Recurrence in Patients Treated with Stereotactic Body Radiotherapy for Early-Stage Non-Small Cell Lung Cancer: A Multicentric Study. J Nucl Med. 2020 Jun; 61(6):814–820. https://doi.org/10.2967/jnumed.119.228106 Epub 2019 Nov 15. PMID: 31732678.

15. Lucia F, Visvikis D, Vallières M, Desseroit MC, Miranda O, Robin P, Bonaffini PA, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2019 Apr; 46(4):864–877. https://doi.org/10.1007/s00259-018-4231-9 Epub 2018 Dec 7. PMID: 30535746.

16. Hatt M, Lucia F, Schick U, Visvikis D. Multicentric validation of radiomics findings:challenges and opportunities. EBioMedicine. 2019 Sep; 47:20–21. https://doi.org/10.1016/j.ebiom.2019.08.054 PMID: 31474549

17. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010 Oct; 49 (7):1012–6. https://doi.org/10.3109/0284186X.2010.498437 PMID: 20831489.

18. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. J Nucl Med. 2015 Nov; 56(11):1667–73. https://doi.org/10.2967/jnumed.115.156927 Epub 2015 Jul 30. PMID: 26229145.

19. Peerlings J, Woodruff HC, Winfield JM, Ibrahim A, Van Beers BE, Heerschap A, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. Sci Rep. 2019 Mar 18; 9(1):4800. https://doi.org/10.1038/s41598-019-41344-5 PMID: 30886309.

20. Da-ano R., Masson I., Lucia F. et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci Rep* 10, 10248 (2020). https://doi.org/10.1038/s41598-020-66110-w PMID: 32581221

21. Hatt M., Parmar C., Qi J. and El Naqa I., "Machine (Deep) Learning Methods for Image Processing and Radiomics," in *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 104–108, March 2019, https://doi.org/10.1109/TRPMS.2019.2899538

22. Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol. 2020 Jul 20. https://doi.org/10.1088/1361-6560/aba798 Epub ahead of print. PMID: 32688357.

23. Chatterjee A. et al., "Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization," in *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 210–215, March 2019, https://doi.org/10.1109/TRPMS.2019.2893860

24. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med. 2018 Aug; 59 (8):1321–1328. https://doi.org/10.2967/jnumed.117.199935 Epub 2018 Jan 4. PMID: 29301932.

25. Stein CK, Qu P, Epstein J, Buros A, Rosenthal A, Crowley J, et al. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC Bioinformatics. 2015 Feb 25; 16:63. https://doi.org/10.1186/s12859-015-0478-3 PMID: 25887219.

26. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, et al. Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinform. 2013 Jul; 14(4):469–90. https://doi.org/10.1093/bib/bbs037 Epub 2012 Jul 31. PMID: 22851511.

27. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 2011 Feb 28; 6 (2):e17238. https://doi.org/10.1371/journal.pone.0017238 PMID: 21386892.

28. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 2010 Aug; 10(4):278–91. https://doi.org/10.1038/tpj.2010.57 PMID: 20676067.

29. Kupfer P, Guthke R, Pohlers D, Huber R, Koczan D, Kinne RW. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. BMC Med Genomics. 2012 Jun 8; 5:23. https://doi.org/10.1186/1755-8794-5-23 PMID: 22682473.

30. Evan Johnson W., Cheng Li, Ariel Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostatistics*, Volume 8, Issue 1, January 2007, Pages 118–127, https://doi.org/10.1093/biostatistics/kxj037 PMID: 16632515

31. Konstantinopoulos P. A., Cannistra S. A., Fountzilas H., Culhane A., Pillay K., Rueda B., et al. (2011). Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer. *PloS one*, 6(3), e18202. https://doi.org/10.1371/journal.pone.0018202 PMID: 21479231

32. Lucia F, Visvikis D, Desseroit MC, Miranda O, Malhaire JP, Robin P, et al. 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. 2018 May; 45(5):768–786. https://doi.org/10.1007/s00259-017-3898-7 Epub 2017 Dec 9. PMID: 29222685.

33. Hatt M, Cheze le Rest C, Turzo A, Roux C, Visvikis D. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. IEEE Trans Med Imaging. 2009 Jun; 28(6):881–93. https://doi.org/10.1109/TMI.2008.2012036 Epub 2009 Jan 13. PMID: 19150782.

34. Pieper S., Halle M. and Kikinis R., "3D Slicer," *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, Arlington, VA, USA, 2004, pp. 632–635 Vol. 1.

35. Zwanenburg A, Leger S, Vallières M, et al., "Image biomarker standardisation initiative-feature definitions," arXiv preprint arXiv:1612.07003, 2016.

36. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJWL, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017 Aug 31; 7(1):10117. https://doi.org/10.1038/s41598-017-10371-5 PMID: 28860628.

37. Witten IH, Frank E, Hall MA, and Pal CJ. *DataMining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

38. Fonti V, and Belitser E. Feature Selection using LASSO. Research Paper in Business Analytics March 2017.

39. Breiman L. "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

40. Vapnik V. N. 1995. The Nature of Statistical Learning Theory. (New York: Springer-Verlag).

41. Varma S, and Simon R. "Bias in error estimation when using cross- validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, p. 91, 2006. https://doi.org/10.1186/1471-2105-7-91 PMID: 16504092

42. Lal TN, Chapelle O, Weston J, Elisseeff A. "Embedded methods" in Feature Extraction: Foundations and Applications Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, pp. 137–165, 2006.

43. Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015 Feb; 42(2):328–54. https://doi.org/10.1007/s00259-014-2961-x Epub 2014 Dec 2. PMID: 25452219.

44. Kaalep A, Sera T, Rijnsdorp S, Yaqub M, Talsma A, Lodge MA, et al. Feasibility of state of the art PET/CT systems performance harmonisation. Eur J Nucl Med Mol Imaging. 2018 Jul; 45(8):1344–1361. https://doi.org/10.1007/s00259-018-3977-4 Epub 2018 Mar 2. PMID: 29500480.

45. Pfaehler E, van Sluis J, Merema BBJ, van Ooijen P, Berendsen RCM, van Velden FHP, et al. Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts. J Nucl Med. 2020 Mar; 61(3):469–476. https://doi.org/10.2967/jnumed.119.229724 Epub 2019 Aug 16. PMID: 31420497.

46. Choe J, Lee SM, Do KH, Lee G, Lee JG, Lee SM, et al. Deep Learning-based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses.

Radiology. 2019 Aug; 292(2):365–373. https://doi.org/10.1148/radiol.2019181960 Epub 2019 Jun 18. PMID: 31210613.

47. Bourbonne V, Jaouen V, Hognon C, Boussion N, Lucia F, Pradier O, et al. Dosimetric Validation of a GAN-Based Pseudo-CT Generation for MRI-Only Stereotactic Brain Radiotherapy. Cancers (Basel). 2021 Mar 3; 13(5):1082. https://doi.org/10.3390/cancers13051082 PMID: 33802499.

48. Dai X, Lei Y, Liu Y, Wang T, Ren L, Curran WJ, et al. Intensity non uniformity correction in MR imaging using residual cycle generative adversarial network. Phys Med Biol. 2020 Nov 27; 65(21):215025. https://doi.org/10.1088/1361-6560/abb31f PMID: 33245059.

49. Cackowski S, Barbier E, Dojat M, and Christen T. comBat versus cycleGAN formulti-center MR images harmonization. Proceedings of Machine Learning Research—Under Review:1–15, 2021. Paper Submission.

50. Hognon C, Tixier F, Gallinato O, et al. Standardization of Multicentric Image Datasets with Generative Adversarial Networks. IEEE MIC 2019.

51. Dai X, Lei Y, Liu Y, Wang T, Ren L, Curran WJ, et al. Intensity non uniformity correction in MR imaging using residual cycle generative adversarial network. Phys Med Biol. 2020 Nov 27; 65(21):215025. https://doi.org/10.1088/1361-6560/abb31f PMID: 33245059.

52. Nakajo M, Jinguji M, Tani A, Kikuno H, Hirahara D, Togami S, et al. Application of a Machine Learning Approach for the Analysis of Clinical and Radiomic Features of Pretreatment [^{18}F]-FDG PET/CT to Predict Prognosis of Patients with Endometrial Cancer. Mol Imaging Biol. 2021 Mar 24. https://doi.org/10.1007/s11307-021-01599-9 Epub ahead of print. PMID: 33763816.

53. Beaumont H, Iannessi A, Bertrand AS, Cucchi JM, Lucidarme O. Harmonization of radiomic feature distributions: impact on classification of hepatic tissue in CT imaging. Eur Radiol. 2021 Jan 18. https://doi.org/10.1007/s00330-020-07641-8 Epub ahead of print. PMID: 33459855.

54. Arendt CT, Leithner D, Mayerhoefer ME, Gibbs P, Czerny C, Arnoldner C, et al. Radiomics of high-resolution computed tomography for the differentiation between cholesteatoma and middle ear inflammation: effects of post-reconstruction methods in a dual-center study. Eur Radiol. 2020 Dec 4. https://doi.org/10.1007/s00330-020-07564-4 Epub ahead of print. PMID: 33277670.

55. Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol. 2021 Mar; 31(3):1460–1470. https://doi.org/10.1007/s00330-020-07174-0 Epub 2020 Sep 9. PMID: 32909055.

56. Wu Q, Wang S, Li L, Wu Q, Qian W, Hu Y, et al. Radiomics Analysis of Computed Tomography helps predict poor prognostic outcome in COVID-19. Theranostics. 2020 Jun 5; 10(16):7231–7244. https://doi.org/10.7150/thno.46428 PMID: 32641989.

57. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu IC, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys. 2018 Jul; 45(7):3449–3459. https://doi.org/10.1002/mp.12967 Epub 2018 Jun 13. Erratum in: Med Phys. 2019 Feb;46(2):1080–1087. PMID: 29763967.

58. Sepehri S.; Tankyevych O.; Upadhaya T.; Visvikis D.; Hatt M.; Cheze Le Rest C. Comparison and Fusion of Machine Learning Algorithms for Prospective Validation of PET/CT Radiomic Features Prognostic Value in Stage II-III Non-Small Cell Lung Cancer. *Diagnostics* 2021, 11, 675. https://doi.org/10.3390/diagnostics11040675 PMID: 33918681

59. Lucia F, Visvikis D, Vallières M, Desseroit MC, Miranda O, Robin P, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2019 Apr; 46(4):864–877. https://doi.org/10.1007/s00259-018-4231-9 Epub 2018 Dec 7. PMID: 30535746.

60. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital—A real life proof of concept. Radiother Oncol. 2016 Dec; 121(3):459–467. https://doi.org/10.1016/j.radonc.2016.10.002 Epub 2016 Oct 28. PMID: 28029405.

61. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clin Cancer Inform. 2020 Mar; 4:184–200. https://doi.org/10.1200/CCI.19.00047 PMID: 32134684.