Journal of Nuclear Medicine, published on September 16, 2021 as doi:10.2967/jnumed.121.262464

A guide to ComBat harmonization of imaging biomarkers in multicenter studies

Fanny Orlhac[1*], Jakoba J. Eertink[2], Anne-Ségolène Cottereau[1,3], Josée M. Zijlstra[2], Catherine Thieblemont[4,5], Michel Meignan[6], Ronald Boellaard[7], Irène Buvat[1]

1: Institut Curie, Université PSL, Inserm, U1288 LITO, Université Paris Saclay, Orsay, France.

2: Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Hematology, Cancer Center Amsterdam, Amsterdam, Netherlands.

3: APHP, Université Paris-Descartes, Hôpital Cochin, Department of Nuclear Medicine, Paris, France.

4: APHP, Hôpital Saint-Louis, Department of hemato-oncology, DMU DHI, Université de Paris, Paris, France.

5: Université de Paris, NF-kappaB, Différenciation et Cancer, Paris, France.

6: APHP, Université Paris-Est, Hôpital Henri Mondor, Lysa Imaging, Créteil, France.

7: Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Amsterdam, Netherlands.

* Corresponding author: Fanny Orlhac, PhD (orlhacf@gmail.com)
U1288-LITO, Institut Curie Centre de Recherche
Bât 101B, rue Henri Becquerel
91401 Orsay cedex, France
Tel: +33 1 56 24 71 99

Word count: 6448 words

Running title: ComBat harmonization guide

ABSTRACT

The impact of PET image acquisition and reconstruction parameters on SUV measurements or radiomic feature values is widely documented. This "scanner" effect is detrimental to the design and validation of predictive or prognostic models and limits the use of large multicenter cohorts. To reduce the impact of this scanner effect, the ComBat method has been proposed and is now used in various contexts. The purpose of this article is to explain and illustrate the use of ComBat based on practical examples. We also give examples in which the ComBat assumptions are not met; thus, ComBat should not be used.

Keywords: Radiomics, Harmonization, Texture Analysis, Multicenter Studies

Noteworthy:

- Guidelines for using the ComBat harmonization method on SUVs, Metabolic Tumor Volume or any radiomic features illustrated with simulated and real data (p 4-8)
- Recommendations on the use of covariates within ComBat (p 9-12)
- Comparison of the ComBat, EARL and z-score harmonization strategies (p 12, 15-16)

INTRODUCTION

The emergence of radiomics in mid-2010 renewed interest in quantitative image analysis for prediction and classification tasks. As radiomics requires large image datasets for designing and validating models, it would largely benefit from pooling images from different sites or from different scanners. However, many quantitative biomarkers and radiomic features are sensitive to a scanner or protocol effect (*1–5*), referred to as the site effect hereafter, underlining the importance of harmonizing image acquisition and reconstruction procedures to reduce multicenter variability before pooling data from different sites. Similarly, when a new radiomic or quantitative image analysis method is developed in one site (site 1), its application to images from a different site (site 2) requires prior verification that the images from the two sites are comparable.

Much effort has been deployed in recent years to propose procedures to harmonize image quality (*6*), including the successful resEARch 4 Life accreditation program (EARL) (*7,8*). However, in retrospective studies, many images have been reconstructed using protocols that did not follow these harmonization guidelines, for which it is impossible to retrieve or perform phantom acquisitions that would be needed to harmonize them a posteriori. Often, the raw data are not stored, hampering any novel reconstruction to target a given image quality. The variability between scans resulting from different acquisition/reconstruction protocols can be reduced using image resampling or filtering (*9,10*), but these techniques require image postprocessing and most often yield a decrease in spatial resolution in the images acquired using the most recent devices, yielding suboptimal image quality for subsequent quantitative and radiomic studies.

To address these site effects, the ComBat harmonization method has been proposed (*11–15*) and has produced satisfactory results in various contexts. Since 2017, at least 51 papers have reported the use of ComBat in radiomic analysis of MRI (36%), CT (34%), or PET images (28%). Of these articles, 41% reported higher performance metrics after ComBat than before, and 41% presented only the results with harmonization. Only 18% of the articles did not report a benefit in using ComBat, without any detrimental effect.

ComBat directly applies to features already extracted from the images without the need for retrieving the images. However, as with any harmonization method, it is based of assumptions that have to be met for the method to generate valid results. The objective of this paper is to explain and demonstrate under which conditions ComBat can be used to harmonize image-derived biomarkers measured in different conditions and when it should be used with caution. We first summarize the theory behind ComBat and then illustrate several use cases to demonstrate its ability to compensate for site effects when properly employed and to answer practical questions a ComBat user might have. We also give examples of situations in which the ComBat assumptions are not met; thus, ComBat should not be used. Finally, we discuss the assets and limitations of ComBat.

All patient data used in the examples below were obtained from previous retrospective studies approved by an institutional review board and the requirement to obtain informed consent was waived.

THEORY OF COMBAT

ComBat was initially introduced in the field of genomics (*16*) and widely used in this field (*17*). ComBat assumes that:

$$y_{ij} = \alpha + \gamma_i + \delta_i \varepsilon_{ij} \tag{1}$$

where j denotes the specific measurement of feature y, i denotes the setting, α corresponds to the average value of the feature of interest y, γ_i is an additive “batch” effect affecting the measurement, δ_i is a multiplicative batch effect and ε_{ij} is an error term. Batch i corresponds to the experimental settings used for making the y measurement, including the possible observer effect, scanner effect or even sample effect.

In medical imaging, y is an image feature (for example, SUV), i denotes the scanner, protocol effect or even observer effect (called the site effect), and j denotes the specific measurement, typically the volume of interest (VOI) in which the measurement is made.

The model therefore assumes that the value of measurement i of a given feature y in VOI j is possibly affected by additive and multiplicative effects that depend on the scanner, protocol or even observer who made the measurement. These effects are common to all measurements $j$of that same quantity y made using the same scanner, protocol or observer. Based on multiple measurements y_{ij} of the same feature y made in VOI j in different images coming from different scanners i, the site effects γ_i and δ_i can be estimated using conditional posterior means (*16*) and subsequently corrected using:

$$y_{ij}^{ComBat} = \frac{y_{ij} - \hat{\alpha} - \hat{\gamma_\imath}}{\hat{\delta_\imath}} + \hat{\alpha} \qquad (2)$$

where $\hat{\alpha}$, $\hat{\gamma_\imath}$ and $\widehat{\delta_\imath}$ are estimators of α, γ_i and $\delta_i$and y_{ij}^{ComBat} is the transformed y_{ij} measurement devoid of the site *i* effect.

ComBat is a data-driven method that does not require any phantom acquisition to estimate the site effect but requires data from the different sites with sufficient sample size. The site effect can be estimated and corrected directly from the available image feature values measured at different sites without having to perform any image processing or any new measurements in the images. ComBat always theoretically improves the alignment of the mean and standard deviation of the distributions given the criterion optimized by the method. A Kolmogorov-Smirnov (KS) test can be used to determine whether the statistical distributions of two sets of feature values are significantly different, in which case ComBat is needed, and to check the effectiveness of the applied transformation. A non-significant KS test suggests that there is no evidence of differences in the two distributions, implying that any subsequent analysis should not be affected by a detectable difference between the distributions.

EXAMPLE

We numerically generated 3000 values drawn from three Gaussian distributions with different means [µ]=8 or 12 or 14 and standard deviations [SD]=3 or 4 or 5 (Table 1), mimicking, for example, SUVmax measured in 3 sets of highly metabolic tumors but with 3 scanners of different generations, of which one had a much higher spatial resolution compared to the others (hence higher SUVmax due to reduced partial volume effect (*18*)). As shown in Figure 1, ComBat can be used in two ways: either to realign the distributions of the three sites to a "virtual" site (*11*), which is neither site A nor site B nor site C, or to realign the data from sites B and C to site A chosen as the “reference” site (or vice versa) (*19*). Contrary to what has been reported (*20*), both approaches lead to the same ranking of the patients, hence identical receiver operating curves for classification tasks, and only the absolute value of the feature changes. Aligning the data to a reference site may be preferable for feature value interpretation, but the reference site selection has no impact on the quality of the realignment. In the following, harmonization will always be performed with respect to a reference site.

COMBAT TO COMPENSATE FOR PROTOCOL DIFFERENCES

The straightforward application of ComBat in medical imaging is to compensate for differences in radiomic feature values obtained from images acquired using different protocols. To illustrate this, we used PET images of 49 lesions from 15 lymphoma patients reconstructed according to the EARL1 and EARL2 standards (*8*), called the “EARL experiment” thereafter. Without harmonization, we observed a systematic deviation between the SUVmax values of the two reconstructions (KS: p-value=0.0002, Figure

2). After applying ComBat considering the EARL2 reconstruction as a reference site, we observed a better concordance of the SUVmax values (p-value=0.6994).

NEED FOR TISSUE-SPECIFIC AND TUMOR-SPECIFIC TRANSFORMATIONS

Since ComBat is a data-driven method, the realignment transformation (Eq 2) is specific to the input data. It is therefore specific to the tissue or tumor type and/or patient population from which it is estimated. For example, in (*12*), the ComBat transformation appropriate for SUVmax was different for liver tissue and breast tumors when pooling 63 patients from site A and 74 patients from site B (Figure 3). In that example, values from site B were realigned to values measured in site A, and the resulting transformations were $SUVmax(A) = 1.05 \times SUVmax(B) + 0.07$ for liver tissue and $SUVmax(A) = 1.13 \times SUVmax(B) + 1.84$ for tumor tissue. This effect of the imaging protocols is different as a function of the structure of interest. SUVmax in the liver is not much impacted by the partial volume effect, as the liver is a large region; hence, it is relatively robust to the difference in spatial resolution in the images produced by the two sites. Therefore, the slope of the transformation was close to 1, and the intercept close to 0. In contrast, the SUVmax in breast tumors is affected by the partial volume effect. This translates into a slope further from 1 and an intercept further from 0. Therefore, unlike what is stated in (*21*), phantom measurements cannot be used to determine the transformations to be applied to feature values measured in one site to convert them to values that would have been obtained at the other site a priori. Given the ComBat assumptions, Eq 2 can only be applied to data affected by the site effect in the same way as the data used to estimate the alpha, gamma and delta parameters of the model. This implies that, for example, a transformation derived for lung tumors should not be applied to lymphoma tumors unless the biomarker of interest is affected by the site effect in the same way in the two tumor types.

NEED FOR A FEATURE-SPECIFIC TRANSFORMATION

Just as transformations are specific to each tissue, they are also specific to each index. For example, using the same data as above (Figure 3), the equations differ for SUVmax ($SUVmax(A) = 1.05 \times SUVmax(B) + 0.07$ for liver tissue) and for the Homogeneity feature ($Homog(A) = 1.06 \times Homog(B) - 0.14$). The transformation has to be estimated for each feature independently because not all features are affected in the same way by the site effect. Some features are relatively immune to the site effect (for example, shape features), unlike others (e.g., SUVmax or metabolic tumor volume).

USING COMBAT TO ADJUST CUTOFF VALUES BETWEEN DIFFERENT SITES

Aligning data from different sites might be extremely useful to adjust the cutoff used to distinguish between groups. Let us take the example of lymphoma patients, for whom it is well known that the total metabolic tumor volume (TMTV) calculated from 18F-FDG PET images is a valuable prognostic factor of progression-free and overall survival (*22*). However, the cutoff value to identify patients with a poor prognosis depends on the segmentation method used for TMTV calculation, and there is no consensus on the optimal segmentation method (*23*). ComBat can thus be used to automatically determine how the cutoff value appropriate for a segmentation method should be shifted to be applicable to TMTV measured using a different segmentation method. To illustrate this, we studied a diffuse large B-cell lymphoma cohort of 280 patients from the REMARC trial (NCT01122472), for whom TMTV was calculated from 18F-FDG PET images using two segmentation methods (*24*). Method 1 (M1) used a threshold of 41% of SUVmax to segment lesions previously identified by a nuclear medicine physician. Method 2 (M2) used a convolutional neural network model (*25*). Using M1, the optimal TMTV cutoff value was 242 mL (T_{M1}) to best distinguish between patients with short and long progression-free survival. Applying that cutoff to TMTV values measured with M2, the Youden Index (YI=Sensitivity+Specificity-1) was 0.18 (Se=41%, Sp=77%, Table 2). Based on TMTV distributions obtained by the two methods (Supplemental Figure 1A-C), ComBat identified the transformation needed to convert M1 TMTV values to values that would have been obtained if M2 segmentation was used: $TMTV_{M2} = 0.61 \times TMTV_{M1} - 28.64$. This formula can be used to determine how the cutoff appropriate for M1 TMTV should be shifted to be applicable to TMTV measured with M2, which was 119 mL (=0.61 x 242 -28.64). With that cutoff value, the YI was 0.22 (Se=64%, Sp=58%), close to the performance obtained when optimizing the cutoff value directly on the M2 TMTV (YI=0.23). These results demonstrate the ability of ComBat to determine how a cutoff should be shifted to account for differences in the segmentation method.

WHEN IS A COVARIATE NEEDED?

Equation (1) above corresponds to the simplest version of ComBat, which is applicable when the two distributions of features to be realigned are drawn from the same population, and only differ because of a site effect. However, in many examples, each of these distributions is itself composed of 2 or more distributions. For example, a feature value distribution might be different in patients with different tumor stages. If the subcategories (patients with different stages) are not present with the same frequencies in the two sites, then the feature distributions observed in the two sites will differ in two respects: because of the site effect and because of the different frequencies of subcategories. Equation 1 will not apply unless the subcategory covariate is introduced. Equation 1 then becomes:

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i \varepsilon_{ij} \tag{3}$$

where X is the design matrix for the covariates of interest, and β is the vector of regression coefficients corresponding to each covariate. The values after realignment are obtained using:

$$y_{ij}^{ComBat} = \frac{y_{ij} - \hat{\alpha} - X_{ij}\hat{\beta} - \hat{\gamma}_i}{\hat{\delta}_i} + \hat{\alpha} \tag{4}$$

To illustrate the impact of using a covariate, we performed 5 experiments, as listed in Table 1 (Experiments 2-6). In all experiments, we assumed we had data from 2 different sites and that in each site, there were patients with “limited” stage or “advanced” stage diseases.

In experiment 2, the numbers of patients with limited-stage and advanced-stage disease were identical at both sites. Using ComBat with or without the stage covariate yields almost identical results (Figure 4). The differences are because only one transformation is estimated without a covariate, compared to two transformations corresponding to each of the two stages in the version including a covariate. As the proportion of patients of each stage is exactly the same, so the stage covariate does not introduce confounding factors. The covariate is thus not necessary, but using it does not influence the ComBat results.

In experiment 3, the samples were the same as in experiment 2, but there were no advanced-stage patients in site B. Without the covariate “stage”, ComBat realigns patients in site A (limited and advanced-stages) with patients in site B (limited-stage only), as shown in Figure 5. Although the realignment of the two distributions seems to be satisfactory, a closer analysis shows that limited-stage patients of site A and site B are not well aligned because ComBat assumed that all site A patients were drawn from a single distribution, identical to that of the site B patients. When stage information is provided as a covariate, the distributions of limited-stage patients from site B are properly realigned with those of limited-stage patients from site A.

The frequency of the covariate may also differ between the two sites, such as in experiment 4 (Table 1). Similar to what was observed for experiment 3, the stage covariate must be introduced in the model to obtain a correct realignment for each stage (Figure 4).

Applying ComBat with a covariate is different from performing ComBat for each subcategory separately. Using a covariate assumes that the site effect is identical for the two (or more) subcategories composing the sample and that only the proportion of individuals in the subcategories differs between the sites. The transformations associated with each subcategory are then constrained to have the same slope and will differ in their intercept only, as the intercept expression includes the design matrix X (Supplemental Figure 2). If that assumption can be made, using ComBat with a covariate should be preferred to performing ComBat independently for each subcategory, as ComBat parameter estimates will benefit from a larger sample. If the site effect is expected to be different for the subcategories (e.g. for different tissue types), then ComBat should be performed for each subcategory independently. However, introducing covariates implies that the transformation will be determined from a smaller number of patients, which may lead to a less reliable estimate. The need for a covariate must therefore be carefully considered.

COMBAT COVARIATES DO NOT INTRODUCE SPURIOUS INFORMATION

Introducing covariates does not artificially add information to the data, as demonstrated by experiment 5 (Figure 4). In that setting, the data were the same as in experiment 4, except that at site B, limited- and advanced-stage patients yielded features with the exact same distribution. When using ComBat with the stage covariate, limited-stage patients from both sites are realigned, advanced-stage patients from both sites are realigned, and the differences in limited and advanced-stage patient feature distributions are reduced after pooling the data from both sites, given that there was a real difference between the 2 stages in site A but not in site B. The stage covariate did not introduce any illegitimate differences between the two stages in patients scanned in site B (Figure 4).

Similarly, when the difference between two categories (here stages) is more detectable on feature values measured in one site (here site B) compared to the other (site A), applying ComBat using a covariate will not corrupt the results (Figure 4). In experiment 6, the gap between the limited and advanced stages is 4 times larger at site B than at site A. After realigning the distributions with ComBat and the stage covariate, the gap between the two stages remains larger in site B (interquartile range (IQR) of feature values from site B after ComBat with covariate=7.5) than in site A (IQR=4.2), thus preserving the original properties of the site B distributions (IQR=8.4) compared to without covariate (IQR=4.7).

The fact that ComBat does not introduce false positives even with the addition of a covariate has been previously demonstrated using sham experiments (15).

The covariate can also take continuous values. In the “EARL experiment”, the addition of the metabolic tumor volume of the VOI in milliliters as a covariate also slightly improved the agreement between the SUVmax obtained with the EARL1 and EARL2 reconstructions (Figure 2). With a reduction in the standard deviation of the Bland-Altman plot from 2.1 SUV to 1.9 SUV.

COMBAT VERSUS A Z-SCORE

Another frequent harmonization method that can be applied a posteriori on feature values is the calculation of z-scores at each site independently (*26*). The feature values at site A are converted into z-scores using the average feature value and associated standard deviation observed over all patients in site A. The same procedure is used for data from site B, using the mean and standard deviation of all measurements made at site B. In doing so, values measured at the 2 sites become comparable. Supplemental Figure 3 shows the result after calculating a z-score from the SUVmax values in the lesions for centers A and B in comparison with Figure 3. Yet, this does not preserve the original range of values, since SUV values vary between -1.5 and 3.6 when expressed in z-scores, against 1.2 SUV and 35.8 SUV on the original data. A second limitation is that it is not possible to account for a covariate. Supplemental Figure 4 shows that the absence of advanced-stage in site B for experiment 3 did not allow the distributions of the limited-stages in the two sites to be aligned correctly when using a z-score, in comparison to Figure 5.

WHEN WILL COMBAT FAIL?

For ComBat to be useful, some basic assumptions must be fulfilled:

1. The distributions of the features to be realigned must be similar except for shift (additive factor) and spread (multiplicative factor) effects. This can be checked by plotting the distributions of the feature values from the 2 sites. ComBat can be used even for non-Gaussian distributions. A log-transformation before applying ComBat (followed by exponentiation after ComBat) can further improve ComBat effectiveness for heavy tailed distributions, as shown in Supplemental Figure 1D.
2. Covariates, if any, that might explain different distributions in the two sites (see point 1 above) have to be identified and taken into account using the design matrix of Equation (3).
3. The different sets of feature values to be realigned have to be independent. If not, it is unlikely that assumption 1 will be met; hence, ComBat will not provide any sound result. A practical example is the realignment of TMTV values as described above but between 2 segmentation methods M1 and M2, where M2 produces the same result as M1 in some examples and produces a different result in others. Unless the cases for which the two methods produce the same segmentation can be predicted and coded as a covariate (for example, in small lesions), ComBat should not be used.

To illustrate the latter, we analyzed TMTV from 140 lymphoma patients. The M1 method corresponds to a threshold set to SUV of 4, and the M2 method corresponds to a majority vote between three segmentation approaches, including the M1 method. In 60 out of 140 cases, M2 led to exactly the same TMTV as M1 and the TMTV was different for all other cases. The TMTV values to be aligned are not independent, which results in a misalignment with ComBat (Supplemental Figure 5), which should realign the cases where the TMTVs are identical and different separately.

4. Determining a single transformation with ComBat from data with tissue or tumor types does not always lead to satisfactory data realignments. This is because different texture patterns are not necessarily affected identically by the image acquisition and reconstruction protocols, so it is not appropriate to realign them all using a single ComBat transformation.

This fully explains why Ibrahim et al (*27*) did not realign the data correctly with ComBat since the 10 phantom patterns in the investigated phantom were affected differently by the pixel spacing. When ComBat was applied separately for each of the textural patterns, the realignments were correct (*28*).

WHICH AMOUNT OF DATA IS NEEDED TO USE COMBAT?

The success of ComBat when only small datasets are available depends on the magnitude of the site effect and on the representativeness of the samples available for each site. In previous studies (*13*), ComBat was successful when the number of patients per site was as low as 20. To illustrate the impact of the number of patients, we re-analyzed data from (*12*) by aligning the feature distribution from site B (74 patients) to site A (63 patients) after estimating the ComBat transformation using only a subset of site B data (74 to 5 patients, 100 repeated random selections). Before ComBat, the distributions from the two sites are different (KS p-value <5%) for SUVmax or Homogeneity measured in the lesions (Supplemental Table 1). After ComBat, the distributions were not significantly different in at least 95 out of 100 tests when the transformation was estimated using 25 patients or more from site B for SUVmax (20 patients for Homogeneity). Supplemental Figure 6 shows the increase in variability in estimating the intercept and the slope of the ComBat transformation when the estimation is based on less and less patients. These results support the recommendation of using ComBat when at least 20 to 30 patients per batch are available.

Note that small sample size to estimate the transformations can also lead to a non-significant KS test because the scanner effect becomes undetectable. In case a covariate is used, a minimum of 20-30 patients per covariate in each batch is also recommended.

A variant of ComBat named B-ComBat that uses a bootstrap approach to determine the parameters of the transformation has been proposed (*20*). However, the use of B-ComBat and the potential benefit of this more computationally demanding approach compared to ComBat have not yet been reported by independent groups.

USING COMBAT IN PRACTICE

Different implementations of ComBat are publicly available (R, Python, MATLAB) and are summarized in Table 3. ComBat can also be used without any third-party software or programming skills using a free online application: https://forlhac.shinyapps.io/Shiny_ComBat/.

DISCUSSION

In this article, we provide a guide to understand and use the ComBat harmonization method correctly. The main advantage of ComBat is that it can be used retrospectively and directly on image features that are already calculated without the need to perform phantom experiments. However, given that ComBat is a data-driven method, a highly recommended practice is to scrutinize the distributions of the feature values from the sites to be pooled before using ComBat. This usually makes it possible to quickly determine whether the assumptions underlying ComBat are fulfilled, especially whether the distributions observed in the different sites are similar except for shift and spread effects. When this is the case, ComBat can be used, otherwise, the reason should first be identified. Often, this is because of the presence of one or more covariate(s), such as patient age, disease stage, treatment, molecular subtype, metabolic volume. When covariates can be identified, it is easy to check if ComBat assumptions are met for each dataset corresponding to a covariate value and whether the site effect impacts the sample corresponding to each covariate identically. If so, ComBat can be used by including that covariate. If the site effect impacts samples corresponding to each covariate differently, then a specific ComBat transformation should be estimated for each sample independently. Examination of feature distributions in tumors can sometimes be challenging, as the variability in the biological signal associated with tumor heterogeneity can hide other sources of variability associated with the site effect. An easy check is to segment a reference region of fixed size in a non-pathological tissue (eg, healthy liver) and observe feature values within that region in images from different sites. This is not sufficient, as it will not give precise information about site effects related to the spatial resolution in the images because the liver usually displays a low-frequency signal. However, we still find it useful to characterize how image quality differs between sites.

ComBat users should keep in mind that data can be grouped in the same batch if they were extracted from images obtained using the same setting on the same scanner. If the image acquisition and reconstruction protocols vary on a scanner, a careful check is needed to ensure that this does not affect the image properties. Otherwise, different batches should be used for the same scanner corresponding to different settings.

In prospective studies, the transformation to be applied with ComBat can be deduced from acquired data previously for the same patient population. The ComBat method is complementary to EARL standardization approach. We have summarized the pros and cons of both approaches in Table 4. EARL and ComBat can be used together if differences in feature distributions remain even with an EARL standardized imaging protocol.

Harmonization in medical imaging can also be seen as domain adaptation, where the goal would be to produce images belonging to a single domain (here, corresponding to the image quality/accuracy obtained with a specific scanner and protocol) from images recorded in different domains. Promising approaches for domain adaptation using, for example, generative adversarial networks (GANs) have been developed in recent years (*29–31*). The role of such approaches in harmonizing PET and SPECT images remains to be studied. Unlike ComBat, GANs act on the images and not on the already computed features, so this requires access to the images, which could be a limitation.

CONCLUSION

In this article, we provide a guide for using the ComBat method to compensate for multicenter effects affecting quantitative biomarkers extracted from nuclear medicine images and beyond. This harmonization method is largely employed in medical imaging and should facilitate large-scale multicenter studies needed to translate radiomics to the clinics.

DISCLOSURE

Michel Meignan has received funds from ROCHE for an educational program. Ronald Boellaard is an (unpaid) scientific advisor for the EARL PET/CT accreditation program. The other authors have no conflicts of interest relevant to this article to declare.

ACKNOWLEDGMENTS

This work is partially supported by the Dutch Cancer Society (# VU 2018-11648). The HOVON-84 study was supported by the Dutch Cancer Society (CKTO-2006-14) and by Roche Nederland. All funding agencies had no role in either the design of the study, the collection, analysis, and interpretation of the data, nor the preparation of the manuscript. We would like to thank the patients and their families, the LYSARC team for the management of the REMARC study, all the investigators, and Celgene / BMS for the financial support.

REFERENCES

1. Yan J, Chu-Shern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. *J Nucl Med*. 2015;56:1667-1673.

2. Reuzé S, Orlhac F, Chargari C, et al. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. *Oncotarget*. 2017;8:43169-43179.

3. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol*. 2017;27:4498-4509.

4. Pfaehler E, van Sluis J, Merema BBJ, et al. Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-dimensionally printed phantom inserts. *J Nucl Med*. 2020;61:469-476.

5. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J Med Imaging*. 2015;2:041002.

6. Clarke LP, Nordstrom RJ, Zhang H, et al. The Quantitative Imaging Network: NCI's historical perspective and planned goals. *Transl Oncol*. 2014;7:1-4.

7. Boellaard R, Delgado-Bolton R, Oyen WJG, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42:328-354.

8. Kaalep A, Sera T, Rijnsdorp S, et al. Feasibility of state of the art PET/CT systems performance harmonisation. *Eur J Nucl Med Mol Imaging*. 2018;45:1344-1361.

9. Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci Rep*. 2018;8:10545.

10. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS ONE*. 2017;12:e0178524.

11. Fortin J-P, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*. 2018;167:104-120.

12. Orlhac F, Boughdad S, Philippe C, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med*. 2018;59:1321-1328.

13. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology*. 2019;291:53-59.

14. Mahon RN, Ghita M, Hugo GD, Weiss E. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Phys Med Biol*. 2020;65:015010.

15. Orlhac F, Lecler A, Savatovski J, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol*. 2021;31:2272-2280.

16. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118-127.

17. Chen C, Grennan K, Badner J, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*. 2011;6:e17238.

18. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. *J Nucl Med*. 2007;48:932-945.

19. Stein CK, Qu P, Epstein J, et al. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics*. 2015;16:63.

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*. 2020;10:10248.

21. Ibrahim A, Primakov S, Beuque M, et al. Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework. *Methods*. 2021;188:20-29.

22. Meignan M, Cottereau AS, Versari A, et al. Baseline metabolic tumor volume predicts outcome in high-tumor-burden follicular lymphoma: a pooled analysis of three multicenter studies. *J Clin Oncol*. 2016;34:3618-3626.

23. Cottereau A-S, Hapdey S, Chartier L, et al. Baseline total metabolic tumor volume measured with fixed or different adaptive thresholding methods equally predicts outcome in peripheral T cell lymphoma. *J Nucl Med*. 2017;58:276-281.

24. Orlhac F, Capobianco N, Cottereau A-S, et al. Refining the stratification of diffuse large B-cell lymphoma patients based on Metabolic Tumor Volume (MTV) by automatically adapting the MTV cut-off value to the segmentation method. *J Nucl Med*. 2020;61:274.

25. Sibille L, Seifert R, Avramovic N, et al. 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*. 2020;294:445-452.

26. Chatterjee A, Vallières M, Dohan A, et al. Creating robust predictive radiomic models for data from independent institutions using normalization. *IEEE Tran Radiat Plasma Med Sci*. 2019;3:210-215.

27. Ibrahim A, Refaee T, Primakov S, et al. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. *Cancers*. 2021;13:1848.

28. Orlhac F, Buvat I. Comment on Ibrahim et al. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. *Cancers*. 2021;13:3037.

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

30. Modanwal G, Vellal A, Buda M, Mazurowski MA. MRI image harmonization using cycle-consistent generative adversarial network. *Medical Imaging 2020: Computer-Aided Diagnosis*. 2020;1131413.

31. Marcadent S, Hofmeister J, Preti MG, Martin SP, Van De Ville D, Montet X. Generative adversarial networks improve the reproducibility and discriminative power of radiomic features. *Radiol Artif Intell*. 2020;2:e190035.

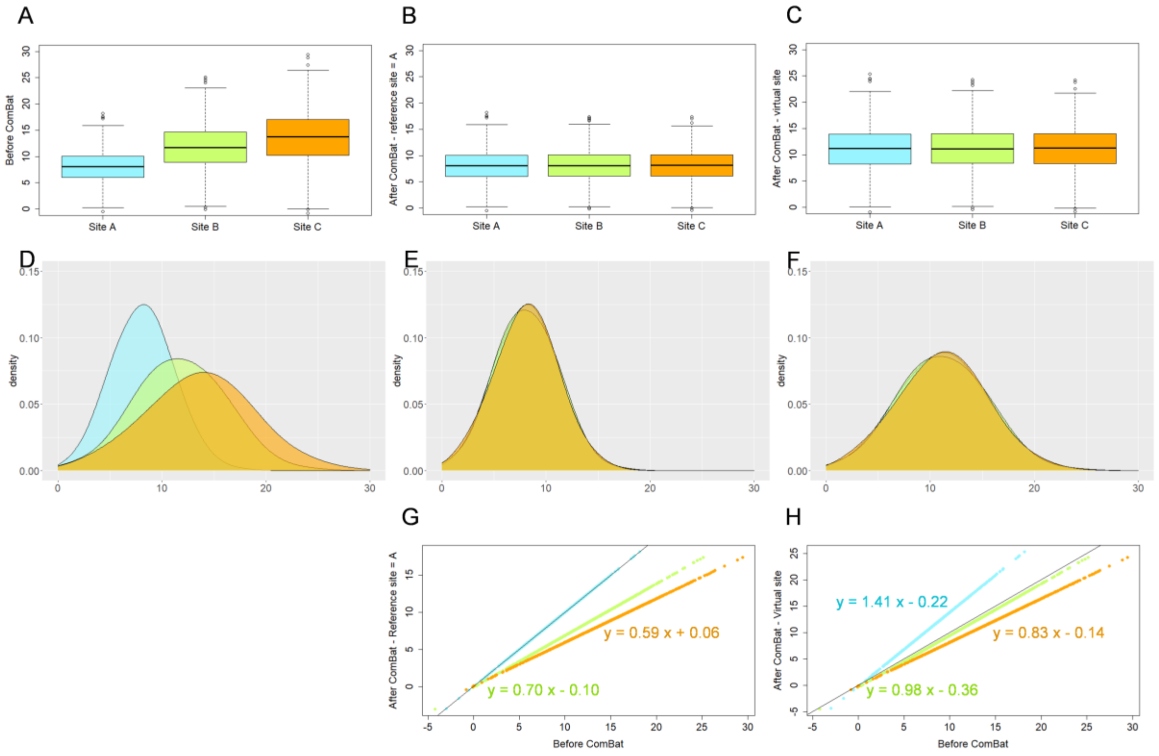


Figure 1: Boxplot and feature value distributions for experiment 1 (details in Table 1). Blue: data for site A, green: data for site B, orange: data for site C. A, D) before ComBat. B, E, G) after ComBat by aligning the data from sites B and C to site A. C, F, H) after ComBat by aligning the data on a "virtual" site (intermediate between the three sites). The bottom graphs show the equations of the transformations.

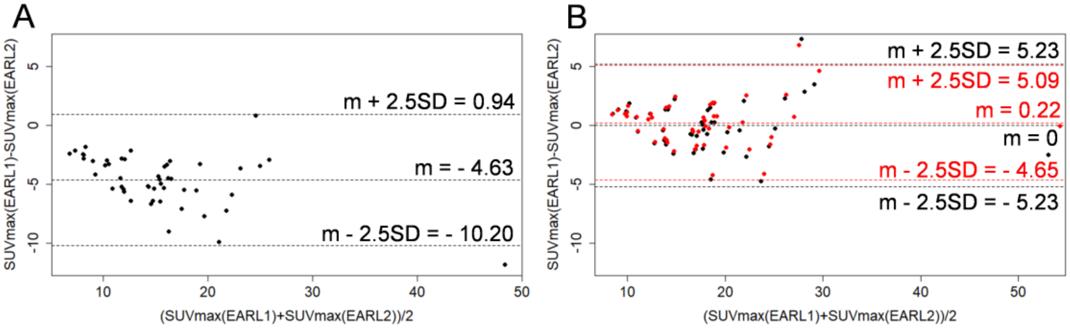


Figure 2: Bland-Altman plots for SUVmax obtained using EARL1 and EARL2 reconstructions. A) before ComBat. B) after ComBat, in black w/o covariate and in red using the metabolic volume (mL) as continuous covariate. m: mean, sd: standard deviation.

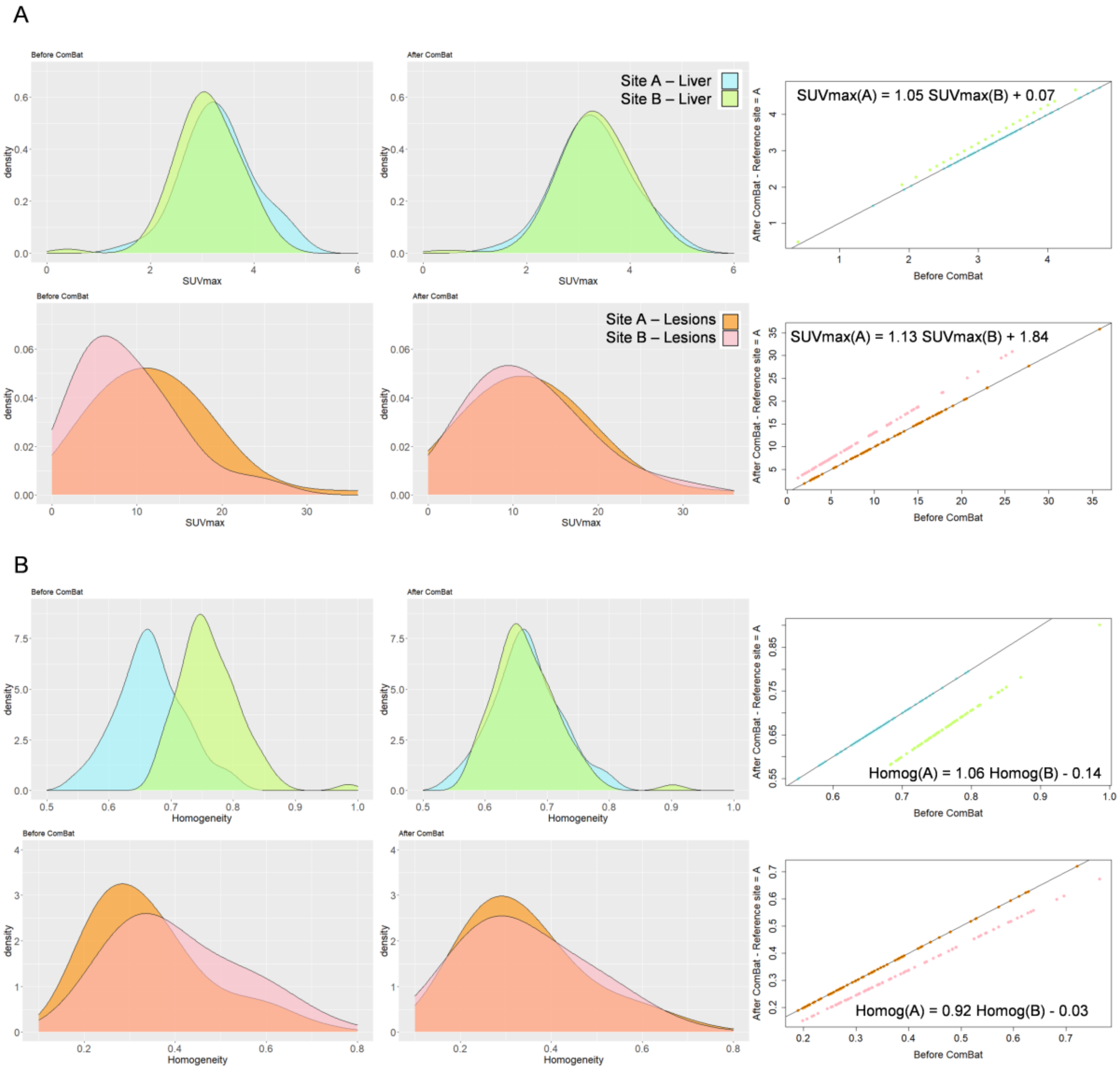


Figure 3: Application of ComBat in the liver (blue and green) and tumor (orange and pink) tissues for A) SUVmax and B) Homogeneity. Left: distributions in the two sites before ComBat. Center: distributions after ComBat (site A: reference site). Right: values after ComBat plotted against value for the same index and tissue before ComBat. The equation is the transformation identified by ComBat to align the data from site B to site A.

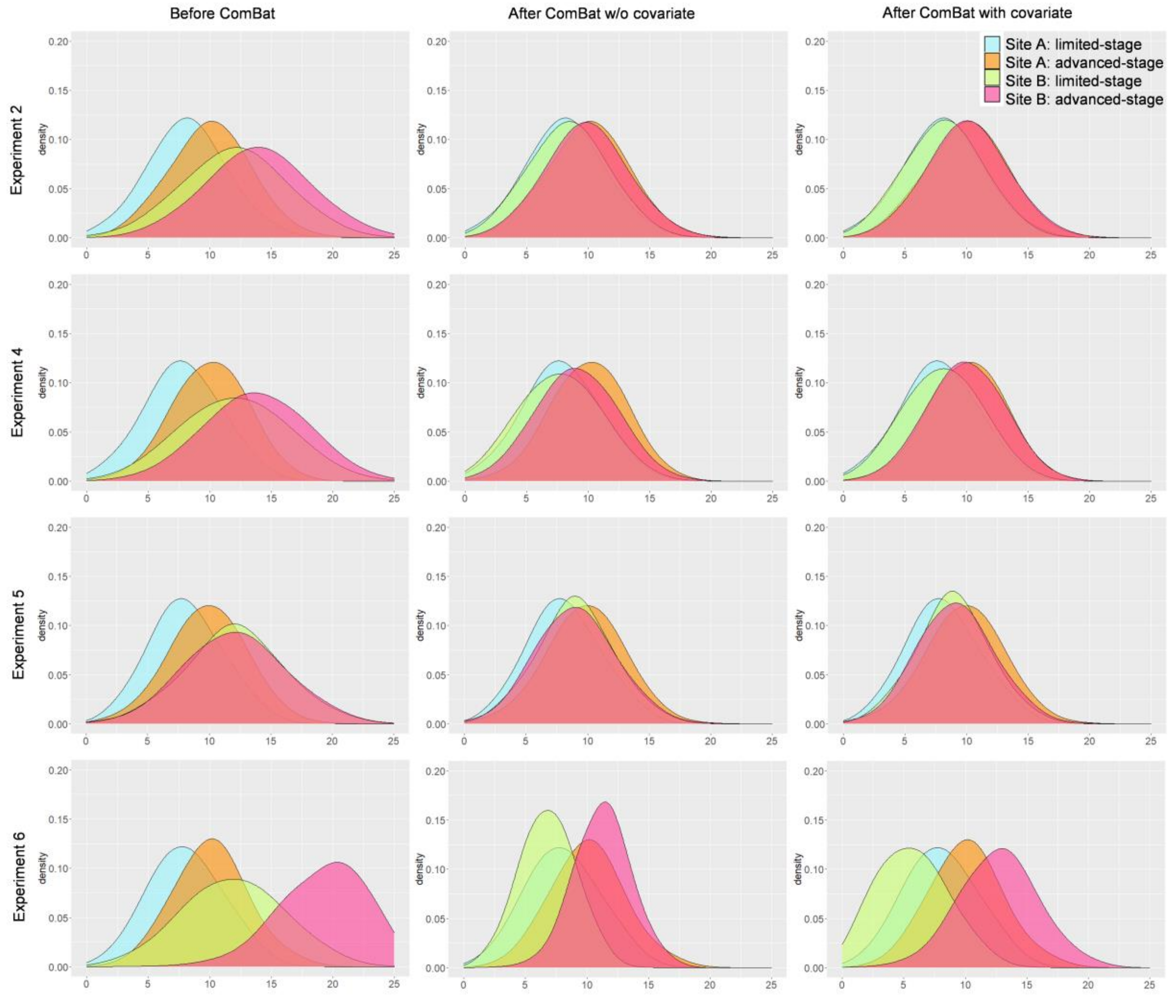


Figure 4: Value distributions for experiments 2, 4, 5 and 6 (details in Table 1). Left: distributions before ComBat. Center: distributions after ComBat (without covariates). Right: distributions after ComBat and specifying stage as a covariate. Blue: data from site A and limited stage. Orange: site A/advanced-stage. Green: site B/limited-stage. Pink: site B/advanced-stage.

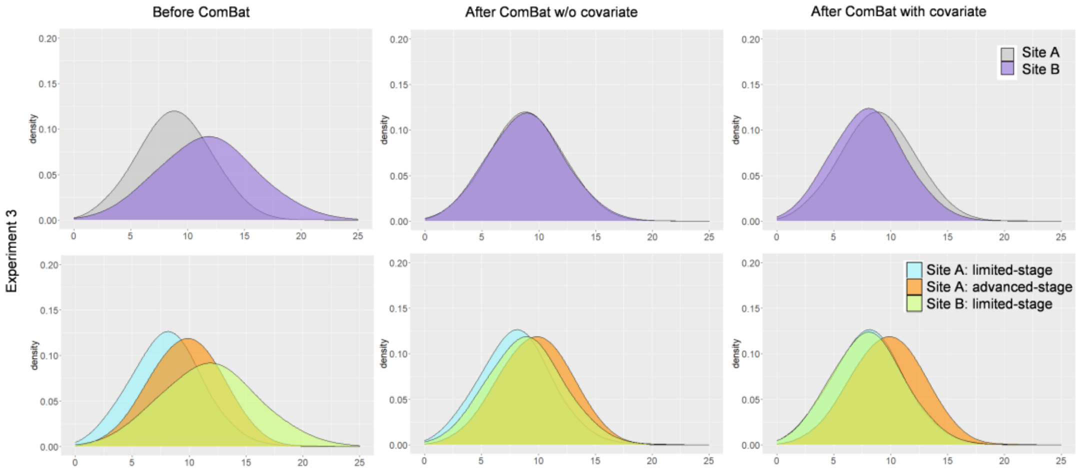


Figure 5: Distributions for experiment 3 (details in Table 1). Left: distributions before ComBat. Center: distributions after ComBat (without covariate). Right: distributions after ComBat and specifying stage as a covariate. Top: by pooling data in each site. Bottom: data represented per site and stage. Blue: data from site A and limited stage. Orange: site A/advanced-stage. Green: site B/limited-stage. Gray: limited and advanced-stages from site A. Purple: limited-stage from site B.

Table 1: Description of the simulations.

	Site A		Site B		Site C
	limited-stage	advanced-stage	limited-stage	advanced-stage	limited-stage
Experiment 1 “Virtual site” Reference site=A	N=1000 µ=8 SD=3	ø	N=1000 µ=12 SD=4	ø	N=1000 µ=14 SD=5
Experiment 2 Reference site=A	N=1000 µ=8 SD=3	N=1000 µ=10 SD=3	N=1000 µ=12 SD=4	N=1000 µ=14 SD=4	ø
Experiment 3 Reference site=A W/o and with covariate (=stage)	N=1000 µ=8 SD=3	N=1000 µ=10 SD=3	N=1000 µ=12 SD=4	ø	ø
Experiment 4 Reference site=A W/o and with covariate (=stage)	N=1000 µ=8 SD=3	N=1000 µ=10 SD=3	N=200 µ=12 SD=4	N=1800 µ=14 SD=4	ø
Experiment 5 Reference site=A W/o and with covariate (=stage)	N=1000 µ=8 SD=3	N=1000 µ=10 SD=3	N=1000 µ=12 SD=4	N=1000 µ=12 SD=4	ø
Experiment 6 Reference site=A W/o and with covariate (=stage)	N=1000 µ=8 SD=3	N=1000 µ=10 SD=3	N=1000 µ=12 SD=4	N=1000 µ=20 SD=4	ø

Note. µ: mean of the Gaussian distribution. SD: standard deviation. N: number of simulated samples. Ø: no simulation for this category.

Table 2: Summary of the results obtained with ComBat to adjust Total Metabolic Tumor Volume cutoff values between different sites.

	Cutoff	Youden	Sensitivity	Specificity
Cut-off optimized for M1	242 mL	0.18	41%	77%
Based on M1 cut-off, estimated cut-off for M2 (ComBat without log-transformation)	119 mL	0.22	64%	58%
Optimal cut-off for M2	112 mL	0.23	66%	57%

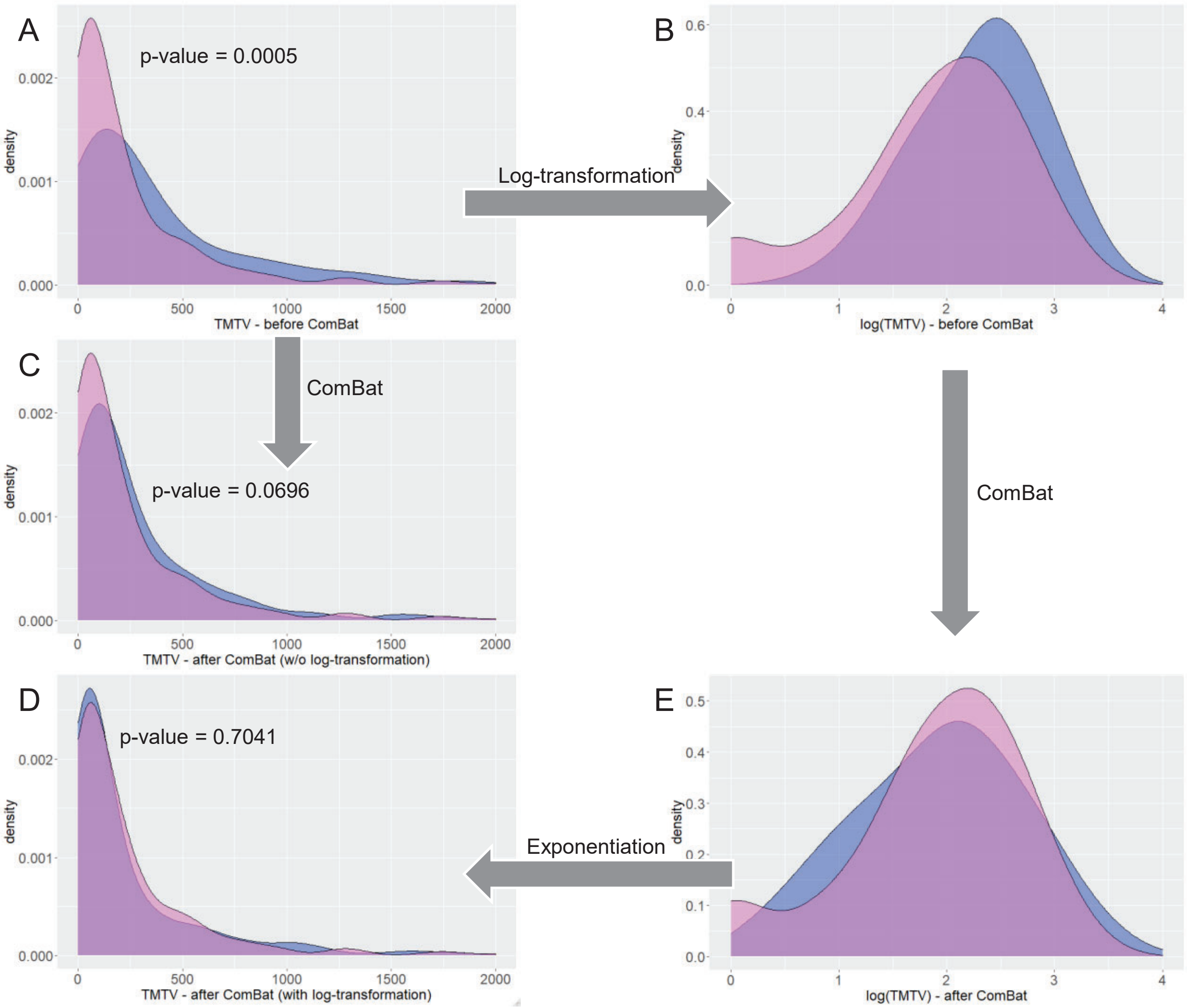
Table 3: Implementations of ComBat.

Name	Details
neuroComBat Script	https://github.com/Jfortin1/ComBatHarmonization Language: R, Python or MATLAB
M-ComBat Script	https://github.com/SteinCK/M-ComBat Language: R
ComBaTool Standalone web application	https://forlhac.shinyapps.io/Shiny_ComBat/ Language: R

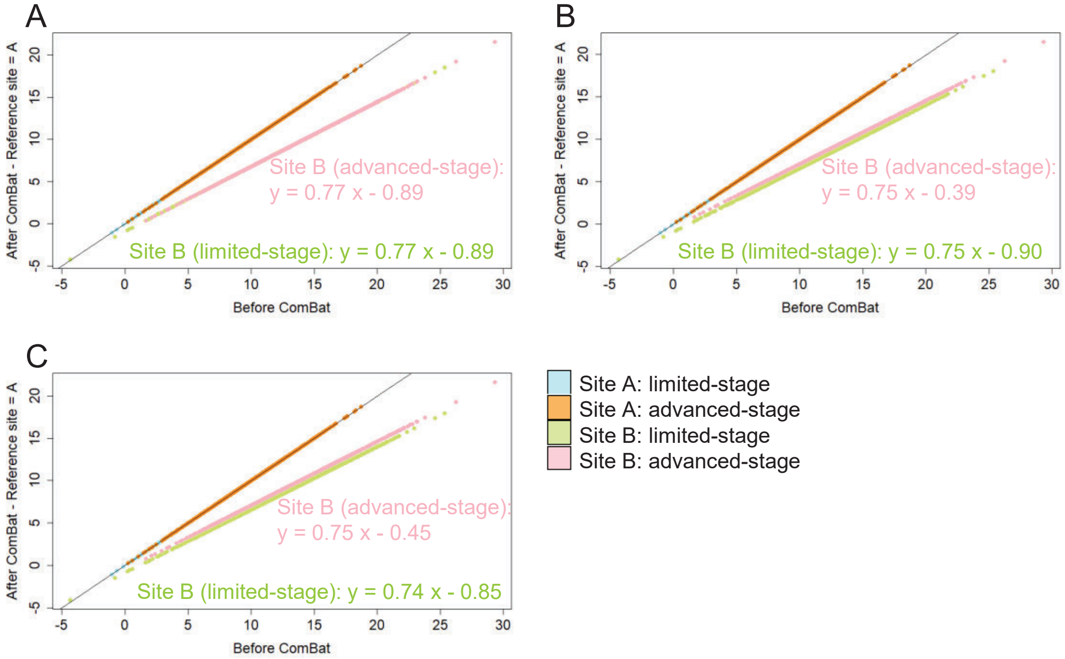
Table 4: Opportunities and limitations of harmonization using EARL and ComBat.

	Upfront harmonization (like EARL)	ComBat
Opportunities	• Applicable without restriction on the number of patients • Valid for any pathology and feature	• Applicable directly on the calculated radiomic feature values (no need to access images) • No need for phantom acquisition • Applicable retrospectively • Applicable prospectively if data have already been acquired for the same pathology with the same acquisition and analysis protocols and settings • Ability to realign data to a particular site
Limitations	• Not applicable retrospectively • Requires acquisition of phantom images, optimization of reconstruction settings and access to the machine	• Minimum number of patients is needed (~20-30 patients per batch) • Specific transformation for each type of tissue, each type of tumor, each scanner, each material in a phantom, each analysis method (e.g., segmentation approach) and each feature • Not applicable prospectively if little or no previously acquired data

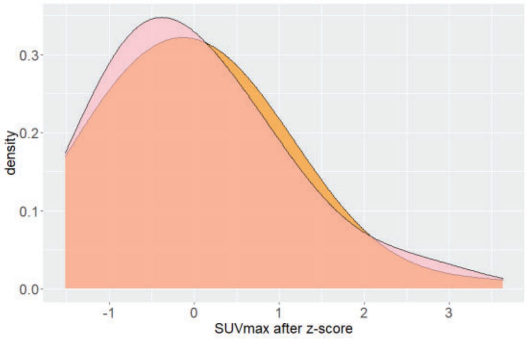
Supplemental Figure 1: Realignment of Total Metabolic Tumor Volume (TMTV) distributions obtained using M1 segmentation method (in blue) on M2 values (in purple) before (A) and after ComBat directly (C) or after a log-transformation of data (B, E, D). P-values are for Kolmogorov-Smirnov tests between the two distributions.



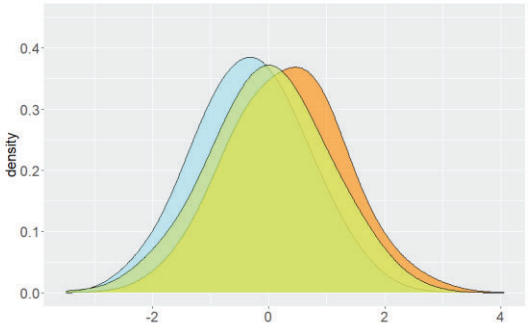
Supplemental Figure 2: Graph of values after ComBat versus values before ComBat for experiment 2 with site A as reference site: A) using ComBat without covariate, B) using ComBat with the stage as covariate, C) using ComBat separately for limited and advanced stages I and II. The equation shown on each graph gives the transformation identified by ComBat to align the data from site B to site A.



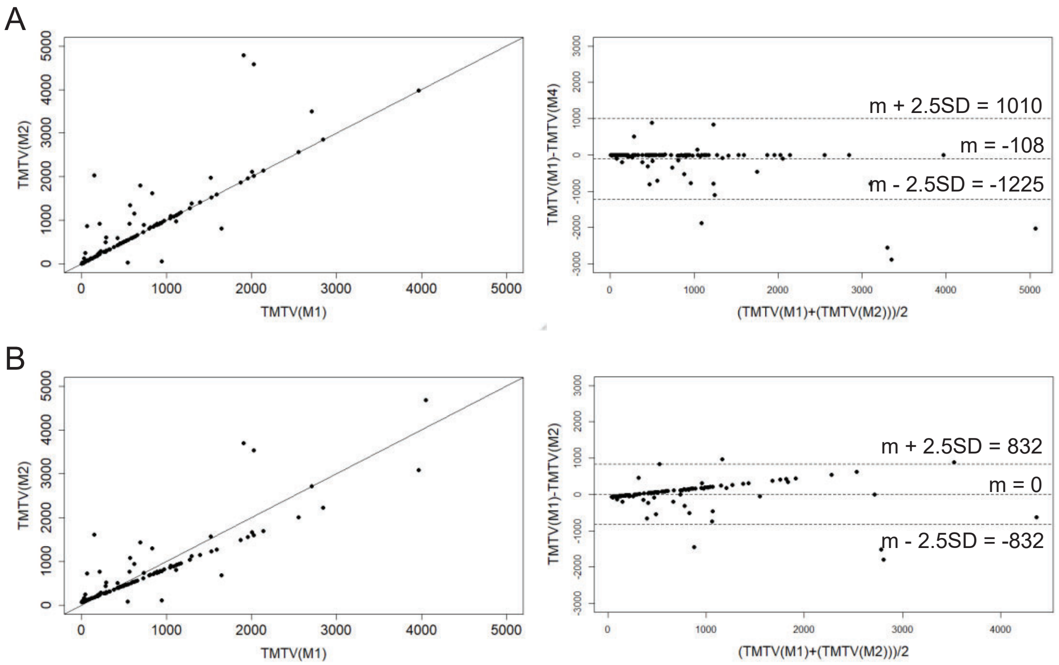
Supplemental Figure 3: Application of z-score for SUVmax values from Figure 3 in tumor tissues. Orange: SUVmax from site A. Pink: SUVmax from site B.



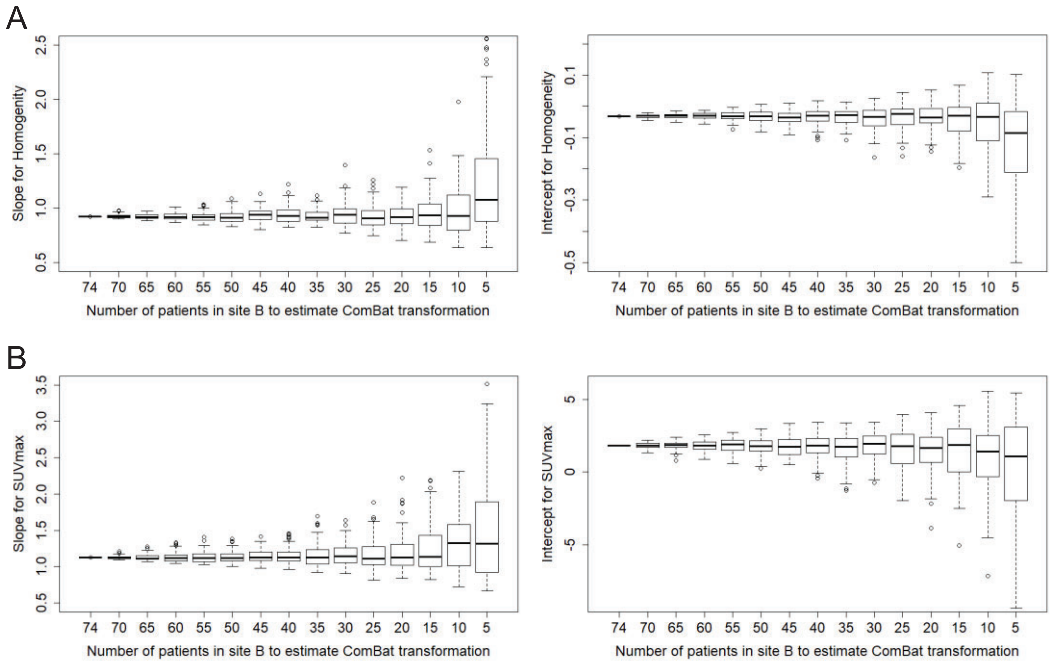
Supplemental Figure 4: Value distributions for experiment 3 (details in Table 1) after z-score calculation in each site separately. Blue: data from site A and limited stage. Orange: site A-advanced stage. Green: site B-limited stage.



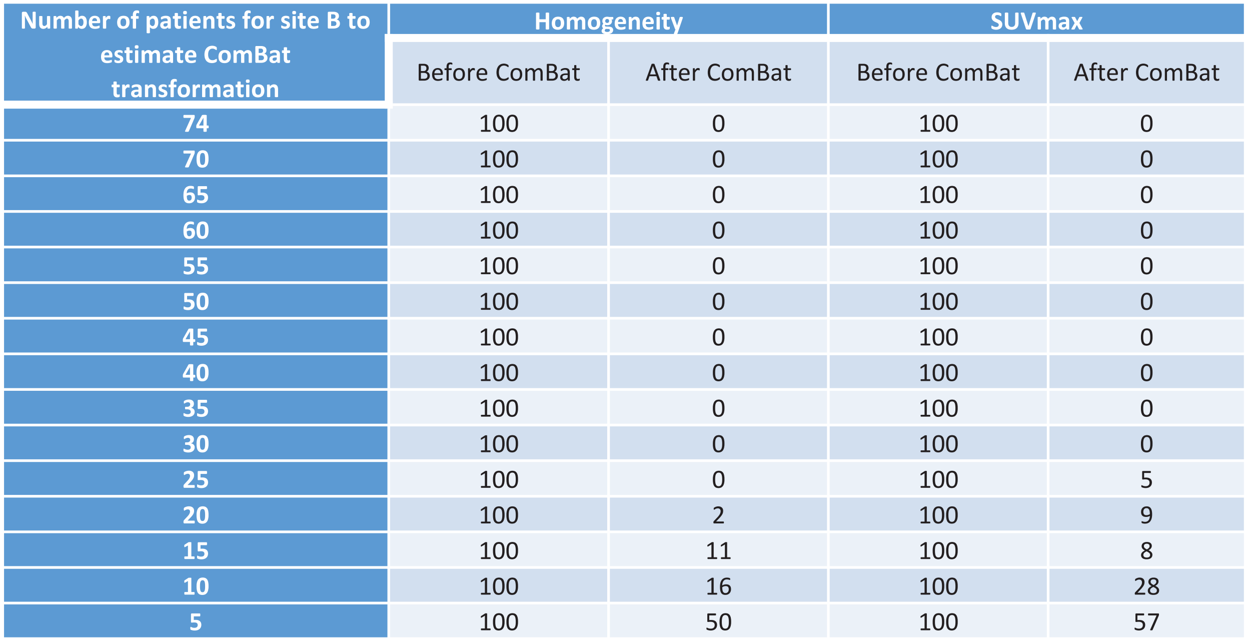
Supplemental Figure 5: Graph of Total Metabolic Tumor Volume (TMTV, on the left) obtained for 140 lymphoma patients using a majority vote (M2) between three segmentation approaches versus one of the methods (M1) and the corresponding Bland-Altman plot (on the right) A) before ComBat and B) after ComBat.



Supplemental Figure 6: Evolution of the slope (left) and intercept (right) of the ComBat transformation from site B to site A for Homogeneity (A) and SUVmax (B) as a function of the number N of patients selected for site B (74 to 5 patients, 100 repeated random selections).



Supplemental Table 1: Number of Kolmogorov-Smirnov tests out of 100 runs with p-value lower than 5% for Homogeneity and SUVmax before and after ComBat. The ComBat transformation from site B to site A is estimated from a subset of patients from site B (from 74 to 5 patients) and then applied to all patients at site B.



Number of patients for site B to estimate ComBat transformation	Homogeneity		SUVmax	
	Before ComBat	After ComBat	Before ComBat	After ComBat
74	100	0	100	0
70	100	0	100	0
65	100	0	100	0
60	100	0	100	0
55	100	0	100	0
50	100	0	100	0
45	100	0	100	0
40	100	0	100	0
35	100	0	100	0
30	100	0	100	0
25	100	0	100	5
20	100	2	100	9
15	100	11	100	8
10	100	16	100	28
5	100	50	100	57