Journal of Nuclear Medicine, published on January 4, 2018 as doi:10.2967/jnumed.117.199935

A post-reconstruction harmonization method for multicenter radiomic studies in PET

Authors: Fanny Orlhac[1*] (PhD), Sarah Boughdad[1,2] (MD), Cathy Philippe[3] (PhD), Hugo Stalla-Bourdillon[3] (MSc), Christophe Nioche[1] (PhD), Laurence Champion[2] (MD), Michaël Soussan[1,4] (MD, PhD), Frédérique Frouin[1] (PhD), Vincent Frouin[3] (PhD), Irène Buvat[1] (PhD)

1: Imagerie Moléculaire *In Vivo*, CEA-SHFJ, Inserm, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France

2: Department of Nuclear Medicine, Institut Curie – René Huguenin, Saint-Cloud, France

3: NeuroSpin/UNATI, CEA, Université Paris-Saclay, Gif-sur-Yvette, France

4: Department of Nuclear Medicine, AP-HP, Hôpital Avicenne, Bobigny, France

*Corresponding author: Fanny Orlhac, PhD
IMIV, CEA-SHFJ, Inserm, CNRS, Univ. Paris-Sud,
Université Paris Saclay
4, place du Général Leclerc
91400 Orsay, France
Tel: 33 1 69 86 78 21
Fax: 33 1 69 86 77 86
Email: orlhacf@gmail.com

Running title: Harmonization for multicenter studies

Word counts: 5974 words

ABSTRACT

Introduction: Several reports have shown that radiomic feature values are affected by the acquisition and reconstruction parameters, thus hampering multicenter studies. We propose a method to standardize features measured from Positron Emission Tomography (PET) images obtained using different imaging protocols to remove the center effect while preserving patient-specific effects.

Methods: Pre-treatment 18F-FDG-PET images of patients with breast cancer were included. In Department A, 63 patients were scanned using a Gemini Time-Of-Flight-PET/Computed-Tomography scanner including 16 triple-negative lesions (TN). In Department B, 74 patients underwent a PET on a GE Discovery 690 Scanner including 15 TN lesions. PET images from Department A were also smoothed using a Gaussian filter to mimic data from a third Department called Department A-S. The primary tumor was segmented to get a tumor volume of interest (VOI) and a spherical VOI was also set in healthy liver tissue. Three Standardized Uptake Values (SUVs) and 6 textural features were computed in all VOI using LIFEx software. A harmonization method, ComBat, initially described for genomic data, was used to estimate the department effect based on the observed feature values. Feature distributions in each department were compared before and after harmonization.

Results: In healthy liver tissue, the feature distributions were significantly different for 4 out of 9 features between Departments A and B, and for 6 out of 9 between Departments A and A-S ($p<0.05$, Wilcoxon's test). After ComBat, none of the 9 feature distributions significantly differ between two departments ($p>0.1$). The same trend was observed in tumors with a realignment of feature values between the departments after ComBat. Identification of TN lesions was largely enhanced after harmonization when the cut-off values were determined on data from one department and applied to data from the other department.

Conclusion: ComBat is efficient at removing the multicenter effect for textural features and SUVs. The method is easy to use, retains biological variations not related to a center effect, and does not require any feature recalculation. Such a harmonization allows for multicenter studies, external validation of radiomic models or cut-off values, and should facilitate the use of radiomic models in clinical practice.

Key words: texture analysis, tumor heterogeneity, PET, radiomics, harmonization

INTRODUCTION

The number of publications involving texture analysis or radiomic studies in medical imaging in general, and in Positron Emission Tomography (PET) in particular, is growing fast (*1,2*). Several authors have published promising results in PET, suggesting that quantification of tumor heterogeneity using radiomic features can predict patient outcome or treatment response (*3–6*). Yet, in a recent study, Chalkidou et al (*7*) pointed an inappropriate control of type I error in many radiomic studies that involved the calculation of a large number of imaging features in a small number of patients. These authors underlined the need for confirming observations and validating models using independent patient cohorts in multicenter settings. Since the first publication including texture analysis in PET images in 2009 (*8*), 77% of radiomic or texture studies in PET included less than 100 patients (Fig. 1), and only three studies involved more than 200 patients (*9–11*). The difficulty in including a large number of patients mostly lies in the need for a clinically homogeneous cohort with respect to the tumor types, stages, treatments, and imaging protocol. Indeed, it has been shown that radiomic feature values are sensitive to the acquisition and reconstruction parameters (*12,13*), thus hindering the pooling of data acquired using different scanners or protocols. More precisely, radiomic feature values are sensitive to the reconstruction algorithm, the number of iterations or subsets, the scan duration per bed position, the post-reconstruction filter and the voxel size (*12–22*). This variability of radiomic feature values implies that a radiomic model established using data from a given PET scanner might not be directly applicable to data from another PET scanner, as recently demonstrated in cervical cancer by Reuzé et al (*23*). This is obviously a severe limitation for the dissemination of radiomic models and their transfer to clinical practice.

In the late 2000's, the genomics field faced a similar problem called "batch effect", where batch refers to the settings used to acquire the data, hence is identical to the scanner or imaging protocol effect in radiomics. In genomics, the batch effect is a technical source of variations caused by the handling of samples (eg, different laboratories, different technicians, different days) potentially masking individual variations, whereas the identification of robust gene signatures to predict disease outcome requires thousands of samples (*24*). Among the methods developed to deal with that batch effect, the ComBat harmonization was described in 2007 (*25*). This method is now widely used in genomics, and has the advantage over other methods that it provides satisfactory results even for small data sets with a limited number of features (*26*).

In this context, the purpose of this study was to determine whether the ComBat method initially described for genomic data analysis could successfully normalize radiomic features as measured in PET so as to remove the center effect while retaining the pathophysiological information, in order to facilitate multicenter studies and exportation of a radiomic model to different centers.

MATERIALS & METHODS

Patients

Two groups of patients with non-metastatic breast cancer were included in this study, with a total of 137 lesions. The first cohort included 63 patients treated at Avicenne Hospital, Bobigny, France, called Department A thereafter. The second cohort consisted of 74 patients scanned at Institut Curie-René Huguenin Hospital, Saint-Cloud, France, called Department B. This study was approved by the local institutional review board (Ile-de-France X), and the requirement to obtain informed consent was waived. A core needle biopsy was performed for all patients to determine the tumor type (*27*). The characteristics of each patient group are summarized in Table 1.

PET/CT imaging protocol

Each 18F-FDG PET/Computed Tomography (CT) scan was performed before the start of therapy. For each patient, capillary blood glucose level was less than 8 mol/mL at the time of 18F-FDG injection.

For patients from Department A, 18F-FDG PET/CT images were acquired using a Gemini Time Of Flight PET/CT scanner (Philips), 78 ± 9 min (range, 59-108) after injection of 18F-FDG (3 MBq/kg). PET images were obtained with 1.45 min per bed position and reconstructed using a list-mode iterative algorithm (Blob-Ordered-Subsets-Time-Of-Flight, 2 iterations, 33 subsets). Attenuation correction was performed using CT images and no post-reconstruction smoothing was used. The PET reconstructed image voxel size was 4 x 4 x 4 mm^3.

In Department B, all 18F-FDG PET/CT images were obtained using a Discovery 690 PET/CT scanner (GE Healthcare), 74 ± 8 min (range, 55-99) after injection of 18F-FDG (3-3.5 MBq/kg) with 2.5 min per bed position. PET images were reconstructed using an Ordered Subset Expectation Maximization iterative reconstruction algorithm (2 iterations, 24 subsets) and a Gaussian post-filtering (full width at half maximum of 6 mm). PET images were corrected for attenuation based on the CT images. The PET reconstructed image voxel size was 2.7 x 2.7 x 3.3 mm^3.

Last, we smoothed PET images from Department A using a 3D-Gaussian filter (sigma = 4 mm) to mimic a third Department called Department A-S.

All PET images were converted in Standardized Uptake Value (SUV) units using standardization by the patient body weight.

Radiomic feature measurements

For each patient, two volumes of interest (VOI) were delineated. First, we segmented the primary lesion using a fixed threshold set to 40% of the maximum SUV in the lesion, called VOI-T thereafter. Second, we located a spherical VOI of about 23 mL in the healthy liver of each patient (VOI-L).

For each VOI, 9 features were measured using LIFEx software (www.lifexsoft.org), including the maximum SUV (SUVmax), the average SUV in the VOI (SUVmean) and the maximum average SUV in a sphere of 1 mL (SUVpeak). For textural feature calculation, voxel intensities were resampled using 64 discrete values between 0 and 20 SUV units, corresponding to an absolute resampling with a bin width about 0.3 SUV (*28*). Six textural features previously selected for their robustness with respect to the segmentation method in each texture correlation group (*29*) were calculated: Homogeneity and Entropy from the co-occurrence matrix, Short-Run Emphasis (SRE) and Long-Run Emphasis (LRE) from the gray-level run length matrix, and High Gray-level Zone Emphasis (HGZE) and Low Gray-level Zone Emphasis (LGZE) from the gray-level zone length matrix. The textural feature calculation method has been described in detail previously (*29*).

Harmonization method

To pool SUV and textural features measured from different PET protocols, we tested a harmonization method previously described for genomic studies to correct the so-called batch effect. The ComBat harmonization model developed by Johnson et al (*25*) assumes that the value of each feature *y* measured in VOI *j* and scanner *i* can be written as:

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i\varepsilon_{ij} \qquad \text{Equation 1}$$

where α is the average value for feature *y*, *X* is a design matrix for the covariates of interest, β is the vector of regression coefficients corresponding to each covariate, γ_i is the additive effect of scanner *i* on features supposed to follow a normal distribution, δ_i describes the multiplicative scanner effect supposed to follow an inverse gamma distribution, and ε_{ij} is an error term (normally distributed with a zero mean), as explained in Fortin et al (*30*). ComBat harmonization consists in estimating γ_i and δ_i using Empirical Bayes estimates (noted γ_i^* and δ_i^*) as described in (*25*). The normalized value of feature *y* for VOI *j* and scanner *i* is then obtained as:

$$y_{ij}^{ComBat} = \frac{y_{ij} - \hat{\alpha} - X_{ij}\hat{\beta} - \gamma_i^*}{\delta_i^*} + \hat{\alpha} + X_{ij}\hat{\beta} \qquad \text{Equation 2}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of parameters α and β respectively. The ComBat harmonization determines a transformation for each feature separately based on the batch (here Department) effect observed on feature values. In the first part of this study, we used ComBat without accounting for any biological covariate (ie X=0), and, in the second part, we used the TN status as the covariate of interest.

For each tissue separately (tumor and liver tissues), we applied ComBat harmonization on all features using the R function called "combat" available at https://github.com/Jfortin1/ComBatHarmonization/.

Statistical analysis

To test the ability of ComBat method to remove the center effect from the feature values, we plotted the probability density function of all features in VOI-L as a function of the department, before and after the ComBat procedure. We used Wilcoxon's tests to determine whether the features were significantly different between departments, and p-values less than 0.05 were interpreted as statistically significant.

For VOI-T, we displayed the boxplots of all features as a function of the tumor type: TN or non-TN lesions, for each department separately, before and after ComBat harmonization. Wilcoxon's tests were used to investigate the impact of ComBat on the feature values in the TN and non-TN groups in each department.

To study the usefulness of ComBat harmonization in multicenter studies, we determined the cut-off value, for each feature, as that maximizing Youden's index defined as (sensitivity + specificity – 1) in separating TN and non-TN groups using data from Department A. We then used these cut-off values to determine whether lesions from Department B were TN or non-TN lesions and we assessed the accuracy of this classification without and with the ComBat harmonization. The accuracy of the TN classification was also measured using the Youden index.

Finally, we investigated the impact of setting a covariate of interest by removing the TN lesions from Department A to get two datasets with different biological compositions: Department A without TN and Department B including 20% of TN. We studied how TN lesions in Department B were distinguished from non-TN lesion in Department A using Wilcoxon's tests before and after ComBat harmonization, without and with TN status as a covariate of interest.

RESULTS

Liver tissue

The plots of feature values in the liver tissue show a shift of the feature distributions between the three departments (Fig. 2; Supplemental Fig. 1). For instance, Homogeneity in VOI-L was lower in Department A than in Departments B and A-S (Fig. 2A). Conversely, SUVmax measured from VOI-L in Department A was higher than in B, which was higher than in A-S (Fig. 2C). Homogeneity, Entropy, SRE and LRE were significantly different between Departments A and B and between Departments A and A-S (p-values of Wilcoxon's test < 0.05, Table 2) when the features were not harmonized. HGZE and SUVmax were also significantly different between Departments A and A-S.

After ComBat harmonization, the distributions better overlapped for all features (Fig. 2; Supplemental Fig. 1) and no feature was significantly different between two departments (p-values>0.1, Table 2).

Tumor tissue

In our cohorts, 16/63 lesions (25%) were TN in Departments A and A-S, and 15/74 lesions (20%) were TN in Department B (Table 1). The mean tumor volume was 28.1±39.1 mL (range: [4.8 - 229.1 mL]) in Departments A and A-S and 12.3±13.0 mL (range: [2.0 – 77.3 mL]) in Department B ($p<0.05$).

In each department separately, Figure 3 and Supplemental Figure 2 show that TN lesions exhibited a higher uptake heterogeneity than non-TN lesions, with lower values for Homogeneity, LRE and LGZE and higher values for Entropy, SRE, HGZE and SUVs.

Before ComBat harmonization, we observed a shift between VOI-T features from the three departments (Supplemental Fig. 2) with, for example, a lower Homogeneity and a higher SUVmax in Department A compared to Departments B and A-S (Fig. 3, Supplemental Fig. 3). Table 3 shows that no feature could distinguish between non-TN lesions from Department A and TN lesions from Department B (p-values>0.08). Five out of 9 features were not significantly different between non-TN lesions from Department A and TN lesions from Department A-S (Supplemental Table 1). All features were significantly different between the two types of lesions in each center separately, except SRE and LRE in Departments A and A-S and Homogeneity in Department A.

After ComBat harmonization, Figure 3 and Supplemental Figure 2 show a realignment of feature values between the three departments for TN lesions (orange, yellow and pink boxes) and for non-TN lesions (green, blue and gray boxes). Table 3 shows that 4/9 features were significantly different between Departments A and B in TN lesions and 8/9 features in non-TN lesions before harmonization. Supplemental Table 1 shows that 7/9 features were significantly different between Departments A and A-S in non-TN lesions before ComBat. After harmonization, feature values were not significantly different between Departments A and B whatever the lesion type (TN or non-TN) (Table 3), except LRE in TN lesions. There were no more significant differences in feature values between Departments A and A-S whatever the lesion type either (Supplemental Table 1, Supplemental Fig. 3). With ComBat harmonization, the p-values of Wilcoxon's tests (Table 3; Supplemental Table 1) were always lower than those obtained without harmonization for distinguishing between non-TN lesions in Department A and TN lesions in Department B or A-S.

To mimic a multicenter study, we determined for each feature a cut-off value for the distinction between TN and non-TN lesions based on data from Department A, and applied these cut-off values on data from Department B. Table 4 shows that, before ComBat harmonization, all Youden indices were between 0.05 and 0.23, which reflects poor to moderate distinction between TN and non-TN lesions in Department B when the cut-off values were established in Department A. After ComBat harmonization, Youden indices of 8/9 features increased and were between 0.20 and 0.36, significantly higher than before harmonization (p-value of paired Wilcoxon's test = 0.008). These Younden index values were close to the ones obtained when the optimal cut-off value was directly determined based on data from Department B (Table 4). For instance, Youden index in Department B for SUVpeak was 0.05 before harmonization when the cut-off was defined in Department A and increased to 0.36 after ComBat harmonization, similar to that obtained when the cut-off was determined directly using the data from Department B (Youden index = 0.37).

Setting a covariate of interest

When we removed the TN lesions from Department A and applied the ComBat harmonization for Departments A and B, the p-values of Wilcoxon's tests for distinguishing between TN lesions in Department B and non-TN lesions in Department A decreased for 8/9 features with respect to the same distinction without ComBat (Fig. 4; Supplemental Fig. 4; Supplemental Table 2). The p-values further decreased when the TN status was set as a covariate of interest in the ComBat harmonization and all features were then statistically different for the distinction between TN lesions in Department B and non-TN lesions in Department A (Supplemental Table 2).

DISCUSSION

In this study, we demonstrate that it is possible to pool radiomic features and SUV measurements from different PET imaging protocols by applying a harmonization method initially used in genomics, ComBat. The efficiency of the method is illustrated using measurements in the healthy liver tissue and in breast lesions in FDG PET images between two different departments and between the same images without or with a 9.4 mm full width at half maximum Gaussian smoothing, ie with significantly different spatial resolutions. In addition, we found that using this method, a cut-off value to distinguish sub-types of lesions established from the data acquired from one PET scanner is applicable to data from another PET scanner.

Among the methods developed to deal with that batch effect, the ComBat method has already been used to normalize histopathological images for cancer diagnosis (*31*) or the cortical thickness measurements from magnetic resonance imaging images (*32*). The ComBat method has several advantages. It is easily accessible, practical to use thanks to an R function available for free, and fast. It is a department-specific harmonization that is based only on patient data acquired in the different departments, and it does not require any phantom experiment, which makes it suitable when analyzing retrospective data. An additional advantage is that it applies directly to the radiomic features and not to the PET images from which the radiomic features are calculated. Therefore, it does not reduce the quality of an image set to match the lower quality of the other sets. Without covariates of interest, affine transformations are used to harmonize the features. The transformations are different for each feature, each VOI type and each Department so that the transformed data lie in a common space in which the department effects have been removed or at least reduced. The transformations are estimated and applied to the measured data themselves so that they can be pooled afterward, without the need for learning sets. The only constraint is the need for data from the different departments so that the transformations can be identified. In the context of radiomic modeling, this implies that if a predictive radiomic model is published based on data acquired in a certain Department A, the model could be applicable to data acquired in a different department if Department A provides the radiomic feature values used to establish the model as well as the model equations and coefficients.

Despite continuous and commendable efforts of the international community and societies to produce guidelines for harmonized imaging procedures (*33–35*), the protocols of acquisition and reconstruction of PET images are not yet standardized. The method we propose offers a solution to perform multicenter studies even when data have been acquired in different conditions. ComBat method is not only usable for radiomic textural features but also for SUV measurements. In our data, we observed that SUVmax was different in the liver between Departments A and B (p=0.05) before harmonization, and that the SUV distributions better overlapped after harmonization (Table 2; Fig. 2). Similarly, in tumors, all SUV p-values were greater than 0.2 between non-TN lesions from Department A and TN lesions from Department B

(Table 3), meaning that SUV measurements failed at distinguishing TN from non-TN lesions. After harmonization, SUV p-values were less than 0.007 between non-TN lesions from Department A and TN lesions from Department B, close to the p-values observed when distinguishing between TN and non-TN lesions in Department A ($p \leq 0.006$) and in Department B ($p \leq 0.02$). This harmonization method is therefore useful when dealing with SUVs in a multicenter investigation or for the retrospective analysis of PET images acquired at the same institution with different scanners or on one scanner but with different acquisition and reconstruction parameters. Supplemental Figure 3 demonstrates the differences in SUV or feature values that can be observed between Departments A-S and A and how they are greatly reduced after ComBat, although some differences remain when looking at individual lesions. Residual differences remaining after ComBat are also due to the strong smoothing in A-S data that induced some information loss, which can obviously not be recovered using ComBat.

In each of the three radiomic studies based on PET images and including more than 200 patients (Fig. 1) to evaluate somatic mutations (*9*) and to predict prognosis (*10*) in non-small cell lung cancers, or to predict treatment response in esophageal cancer (*11*), PET images came from different PET scanners but the scanner effect was not explicitly accounted for. Many studies have reported the impact of acquisition and reconstruction parameters on radiomic feature values. Ignoring the scanner effect when pooling data from different centers can affect the results in two ways: it can either make the results more significant than they actually are or it can hide significant differences (Fig. 5).

Despite the fact that Entropy was reported as the most robust feature in previous publications (Supplemental Table 4 of (*22*)), we observed a shift in entropy values in the liver tissue between the three departments (Supplemental Fig. 1A) with higher Entropy values in Department A than in Department B or A-S (p-value < 0.0001, Table 2). After harmonization, the shift was removed (p-value of Wilcoxon's test > 0.7). This suggests that, even when a feature is identified as robust with respect to different imaging protocols, a scanner effect can still be present and require compensation in multicenter studies.

As explained in Fortin et al (*32*), the ComBat method may be used even if the patient groups have different characteristics by properly setting covariates of interest (equation 1). The harmonization procedure therefore removes the center effect without altering the biological information conveyed by the radiomic features. When patient characteristics are very different between departments, ComBat harmonization should include the definition of covariates of interest that get protected, ie that will not enter the harmonization process. For example, when removing the TN lesions from Department A, we observed better performance for the discrimination between TN lesions from Department B and non-TN lesions from Department A (Supplemental Table 2) when the TN status was included as a covariate of interest using the

X design matrix (equation 1) than when no covariate was defined. Yet, ComBat being a data-driven technique, it is preferable to use clinically and biologically similar datasets whenever possible.

We used the non-parametric version of ComBat. A parametric version of ComBat including assumptions regarding the statistical distribution of the model parameters (equation 1) has also been described (*25*). As these assumptions were not closely fulfilled in our data, we used the non-parametric model instead.

Differences between images from different PET systems could also be due to calibration differences. We checked that simple corrections involving a rescaling or offset factor estimated from a healthy region (healthy liver) were not sufficient to remove the differences between images while ComBat harmonization was successful (data not shown).

A limitation of our study was the limited number of patients in each department, although consistent with most PET radiomic studies (Fig. 1). Further studies are needed to more extensively validate the use of ComBat harmonization for other imaging protocols and other cancer types in the context of radiomics. The minimum number of patients from each department required to use ComBat should be further explored, especially in the non-parametric setting. ComBat has been specifically designed to be robust to small samples (*25, 32)* and has been used in genomics with as few as 25 samples in each batch (*25*).

CONCLUSION

Using the ComBat harmonization procedure initially described for genomic analysis, we showed that radiomic feature values and SUV measurements derived from images acquired in different departments or under different conditions could be pooled for further analysis. The individual variations are preserved in healthy liver tissue and breast lesions after harmonization while the imaging protocol effect is removed. This method is easily available and does not require any feature re-calculation since it directly applies to the radiomic feature values as opposed to the images. The ComBat approach appears to be a promising procedure to build radiomic models from data pooled from different departments and to implement a radiomic model derived from data acquired using a certain PET imaging protocol to data acquired using a different protocol.

ACKNOWLEDGMENT

This work is supported by the "Lidex-PIM" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

FINANCIAL DISCLOSURE

The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

REFERENCES

1. Alic L, Niessen WJ, Veenland JF. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. *PLoS One*. 2014;9:e110300.

2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278:563–577.

3. Ha S, Park S, Bang J-I, Kim E-K, Lee H-Y. Metabolic radiomics for pretreatment (18)F-FDG PET/CT to characterize locally advanced breast cancer: histopathologic characteristics, response to neoadjuvant chemotherapy, and prognosis. *Sci Rep*. 2017;7:1556.

4. Ben Bouallègue F, Al Tabaa Y, Kafrouni M, Cartron G, Vauchot F, Mariano-Goulart D. Association between textural and morphological tumor indices on baseline PET-CT and early metabolic response on interim PET-CT in bulky malignant lymphomas. *Med Phys*. 2017;44:4608-4619.

5. Desbordes P, Ruan S, Modzelewski R, et al. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. *PLoS One*. 2017;12:e0173208.

6. Beukinga RJ, Hulshoff JB, van Dijk LV, et al. Predicting response to neoadjuvant chemoradiotherapy in esophageal cancer with textural features derived from pretreatment (18)F-FDG PET/CT imaging. *J Nucl Med*. 2017;58:723–729.

7. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One*. 2015;10:e0124165.

8. El Naqa I, Grigsby P, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit*. 2009;42:1162–1171.

9. Yip SSF, Kim J, Coroller TP, et al. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. *J Nucl Med*. 2017;58:569–576.

10. Ohri N, Duan F, Snyder BS, et al. Pretreatment 18F-FDG PET textural features in locally advanced non-small cell lung cancer: secondary analysis of ACRIN 6668/RTOG 0235. *J Nucl Med*. 2016;57:842–848.

11. van Rossum PSN, Fried DV, Zhang L, et al. The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer. *J Nucl Med*. 2016;57:691–700.

12. 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;49:1012–1016.

13. 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.

14. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in (18)F-FDG-PET scans of oesophageal cancer. *Eur Radiol*. 2015;25:2805–2812.

15. 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.

16. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. *Mol Imaging Biol*. 2016;18:788–795.

17. Bailly C, Bodet-Milin C, Couespel S, et al. Revisiting the robustness of PET-based textural features in the context of multi-centric trials. *PLoS One*. 2016;11:e0159984.

18. Forgacs A, Pall Jonsson H, Dahlbom M, et al. A study on the basic criteria for selecting heterogeneity parameters of F18-FDG PET images. *PLoS One*. 2016;11:e0164113.

19. Lasnon C, Majdoub M, Lavigne B, et al. (18)F-FDG PET/CT heterogeneity quantification through textural features in the era of harmonisation programs: a focus on lung cancer. *Eur J Nucl Med Mol Imaging*. 2016;43:2324–2335.

20. Orlhac F, Thézé B, Soussan M, Boisgard R, Buvat I. Multi-scale texture analysis: from 18F-FDG PET images to pathological slides. *J Nucl Med*. 2016;57:1823-1828.

21. 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 [Epub ahead of print].

22. Orlhac F, Nioche C, Soussan M, Buvat I. Understanding changes in tumor texture indices in PET: a comparison between visual assessment and index values in simulated and patient data. *J Nucl Med*. 2017;58:387–392.

23. 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;8:43169–433179.

24. Lazar C, Meganck S, Taminau J, et al. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform*. 2013;14:469–490.

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

26. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol*. 2017;35:498–507.

27. Soussan M, Orlhac F, Boubaya M, et al. Relationship between tumor heterogeneity measured on FDG-PET/CT and pathological prognostic factors in invasive breast cancer. *PLoS One*. 2014;9:e94017.

28. Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. *PLoS One*. 2015;10:e0145063.

29. Orlhac F, Soussan M, Maisonobe J-A, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J Nucl Med*. 2014;55:414–422.

30. Fortin J-P, Parker D, Tunc B, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*. 2017;161:149-170.

31. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, Wang MD. Removing batch effects from histopathological images for enhanced cancer diagnosis. *IEEE J Biomed Health Inform*. 2014;18:765–772.

32. Fortin J-P, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *bioRxiv*. 2017;148502.

33. Delbeke D, Coleman RE, Guiberteau MJ, et al. Procedure guideline for tumor imaging with 18F-FDG PET/CT 1.0. *J Nucl Med*. 2006;47:885–895.

34. 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.

35. 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.

Figure 1: Number of patients involved in texture or radiomic studies from PET images since 2009: search from PubMed with the key words "(radiomics OR texture OR textural) AND PET".

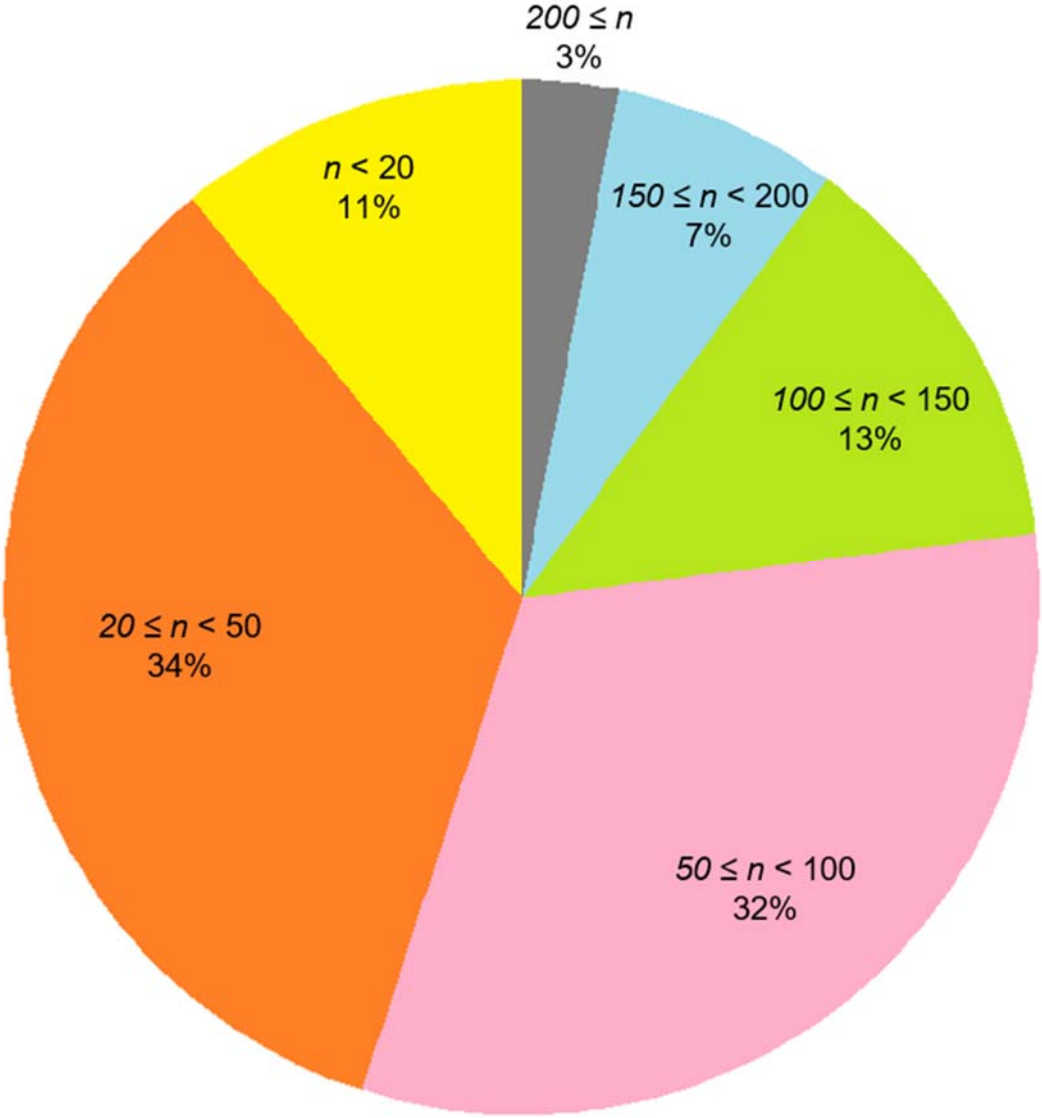


Figure 2: Probability density function (%) of Homogeneity (A, B) and SUVmax (C, D) in liver tissue as observed in Department A (pink), Department B (green) and Department A-S (blue), before (left) and after ComBat harmonization (right).

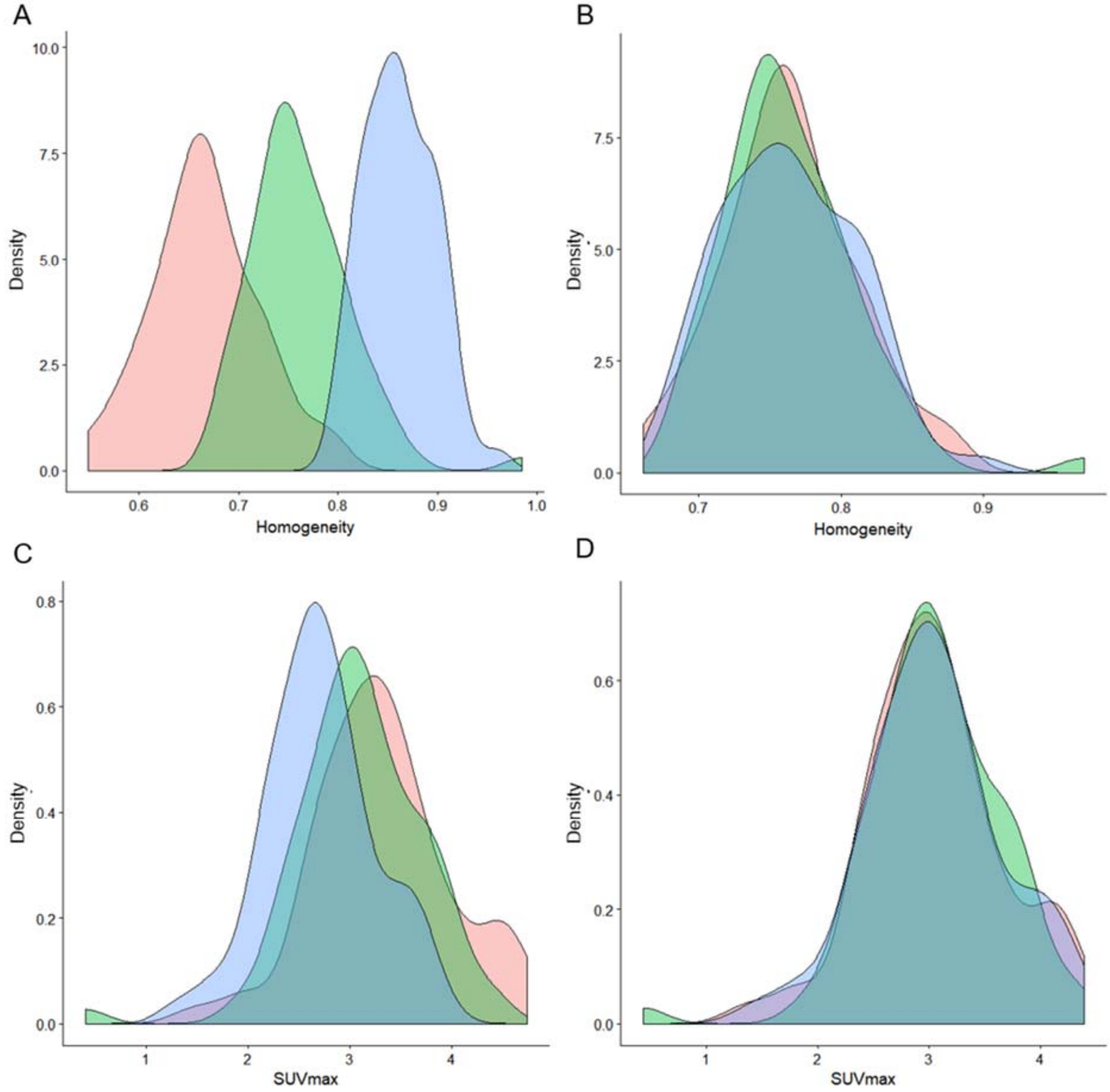


Figure 3: Boxplots of Homogeneity (A) and SUVmax (B) for triple-negative (TN) and non-triple-negative (non-TN) lesions before and after ComBat harmonization in the three departments separately.

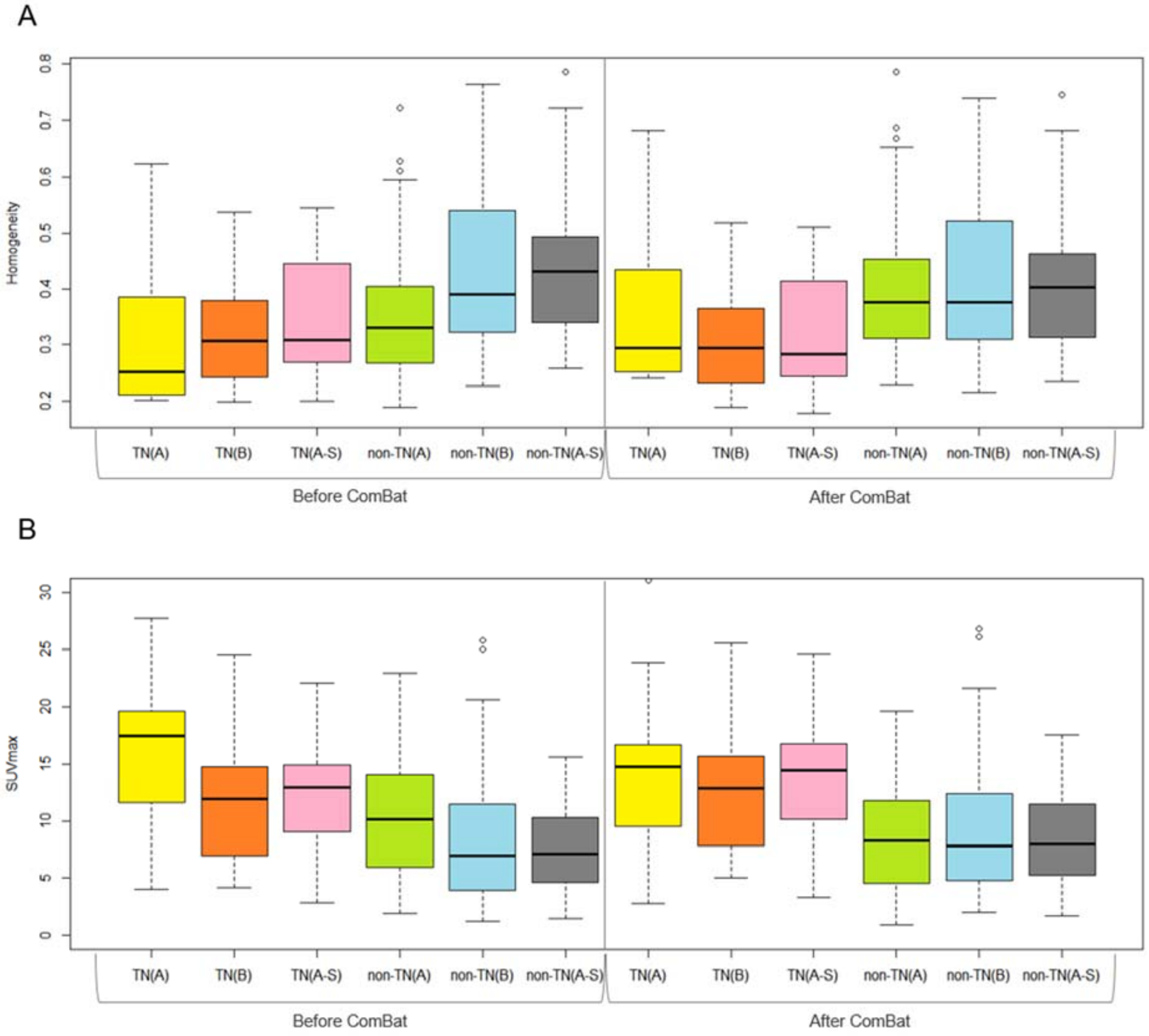


Figure 4: Boxplots of Homogeneity (A) and SUVmax (B) for triple-negative (TN) and non-triple-negative (non-TN) lesions before and after ComBat harmonization without and with TN status as a covariate, for Departments A and B separately when removing all TN in Department A to determine the transformations.

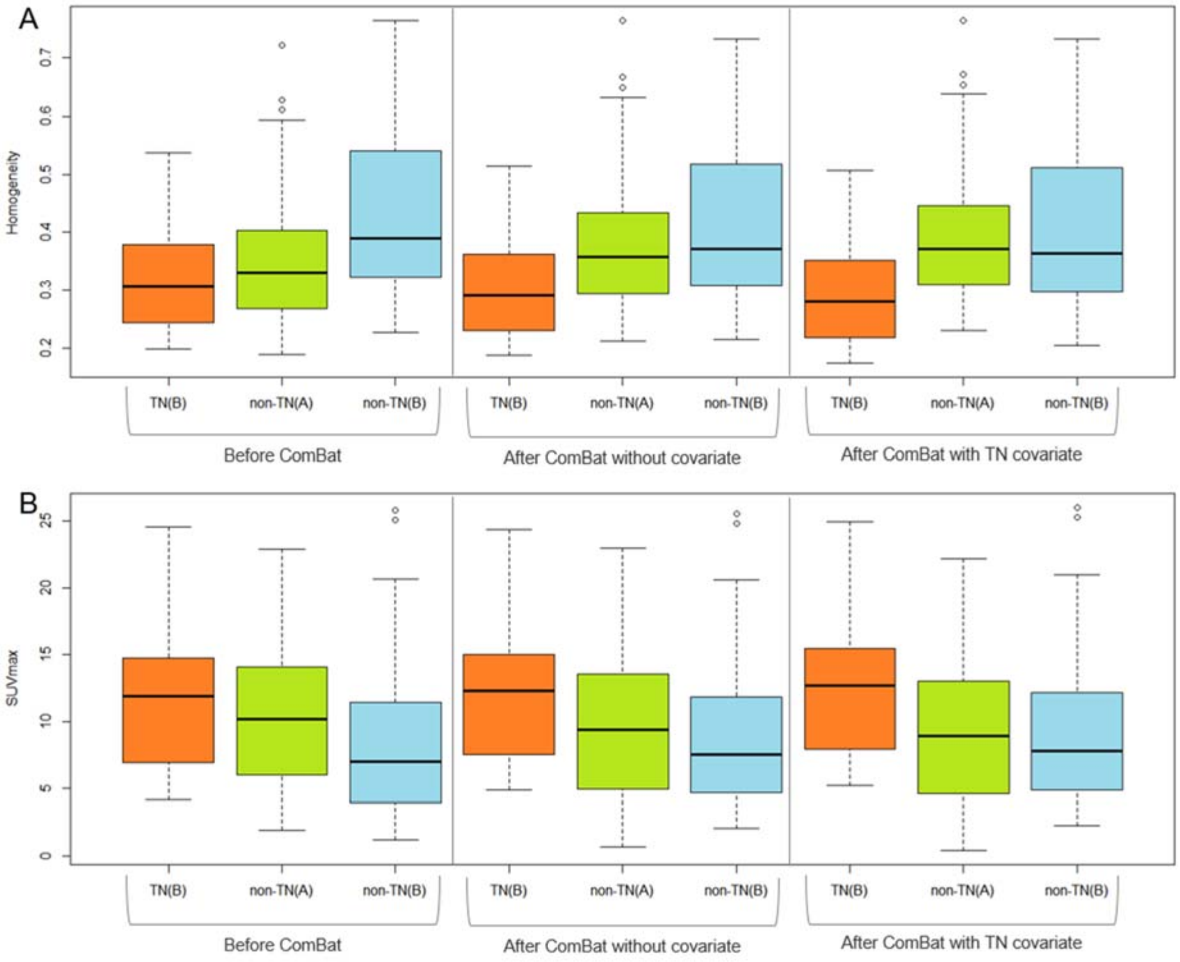
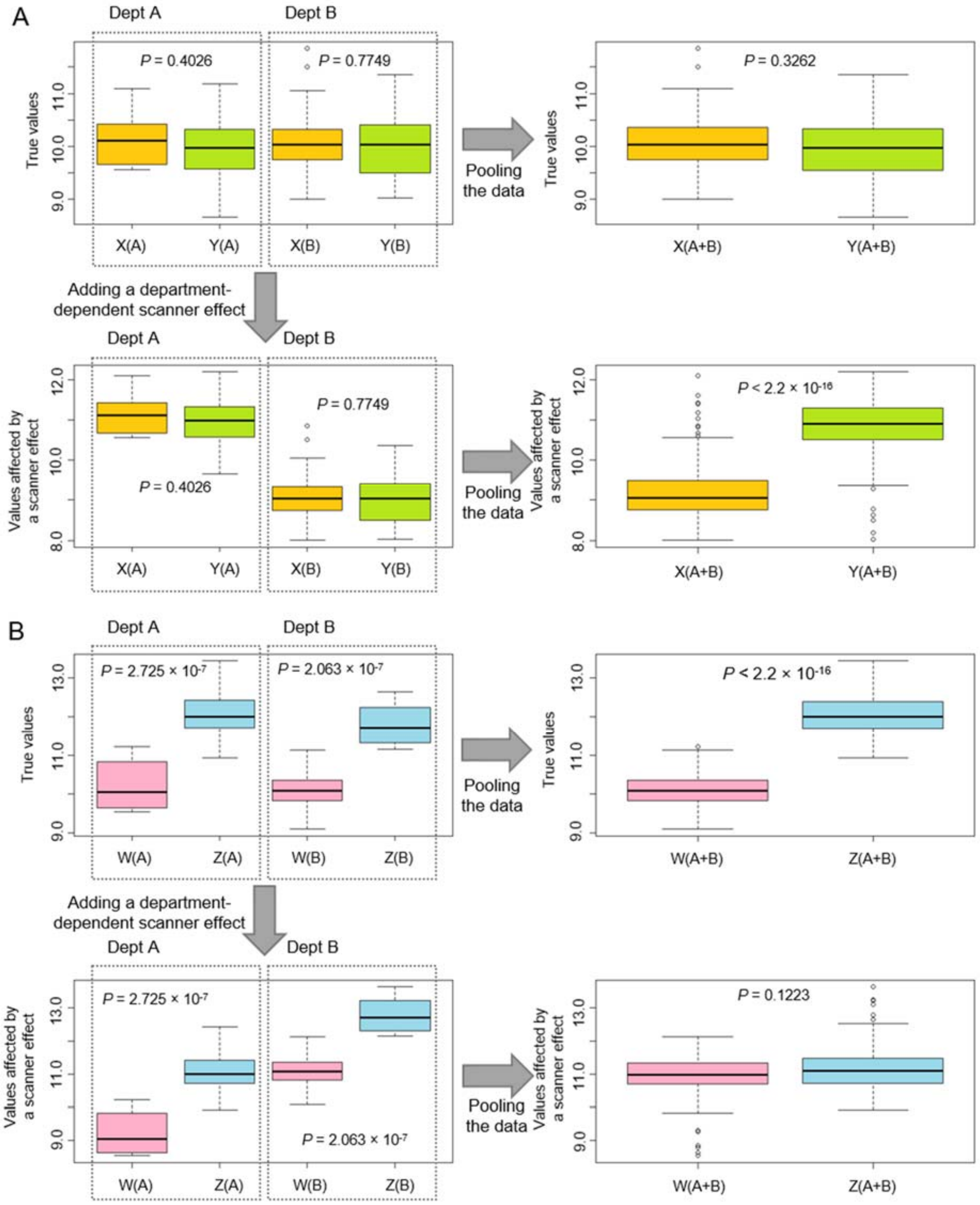


Figure 5: Simulations of results obtained when pooling data from different imaging protocols.

A. True values for two hypothetical types of tumors (X and Y) in two departments (A and B) randomly drawn from normal distributions with the same mean (m=10) and standard deviation (sd=0.5) with 10 X(A) "tumors", 10 Y(B) "tumors", 100 Y(A) "tumors" and 100 X(B) "tumors". When pooling the data from the 2 departments (top row, right), there was no significant difference between the two types of lesions (Wilcoxon's test). When modeling a department-dependent scanner effect (+1 for department A and -1 for department B), the two types of lesions were statistically different due to the scanner effect only.

B. True values for two hypothetical types of lesions W (normal distribution, mean=10, sd=0.5) and Z (normal distribution, mean=12, sd=0.5) significantly different in each department, with 10 W(A) "tumors", 10 Z(B) "tumors", 100 Z(A) "tumors" and 100 W(B) "tumors". When pooling the data, we observed a significant difference between the two types of lesions. When adding a department-dependent scanner effect (-1 for department A and +1 for department B), the measurements in the two types of lesions were no longer statistically different.


A
Dept A
Dept B
$P = 0.4026$
$P = 0.7749$
True values
9.0
10.0
11.0
X(A)
Y(A)
X(B)
Y(B)
Pooling the data
$P = 0.3262$
X(A+B)
Y(A+B)
Adding a department-dependent scanner effect
Dept A
Dept B
Values affected by a scanner effect
8.0
10.0
12.0
$P = 0.7749$
$P = 0.4026$
X(A)
Y(A)
X(B)
Y(B)
Pooling the data
$P < 2.2 \times 10^{-16}$
X(A+B)
Y(A+B)
B
Dept A
Dept B
$P = 2.725 \times 10^{-7}$
$P = 2.063 \times 10^{-7}$
True values
9.0
11.0
13.0
W(A)
Z(A)
W(B)
Z(B)
Pooling the data
$P < 2.2 \times 10^{-16}$
W(A+B)
Z(A+B)
Adding a department-dependent scanner effect
Dept A
Dept B
Values affected by a scanner effect
9.0
11.0
13.0
$P = 2.725 \times 10^{-7}$
$P = 2.063 \times 10^{-7}$
W(A)
Z(A)
W(B)
Z(B)
Pooling the data
$P = 0.1223$
W(A+B)
Z(A+B)

	Department A	Department B
Age (mean±sd)	55±15	51±14
Molecular subtypes		
Luminal A	9 (14%)	11 (15%)
Luminal B	35 (56%)	44 (59%)
Triple-negative (TN)	16 (25%)	15 (20%)
Her2+	3 (5%)	2 (3%)
Unknown	-	2(3%)

Table 1: Patient characteristics.

	Department A vs Department B		Department A vs Department A-S	
	Before Combat	**After ComBat**	**Before Combat**	**After ComBat**
Homogeneity	**<0.0001**	0.7592	**<0.0001**	0.9300
Entropy	**<0.0001**	0.7828	**<0.0001**	0.9611
SRE	**<0.0001**	0.8930	**<0.0001**	0.7922
LRE	**<0.0001**	0.4708	**<0.0001**	0.8491
LGZE	0.5961	0.1319	0.9397	0.9650
HGZE	0.2328	0.8100	**0.0233**	0.8759
SUVmax	0.0522	0.7424	**<0.0001**	1.0000
SUVmean	0.4042	0.8409	0.9980	1.0000
SUVpeak	0.3407	0.9666	0.0614	0.9766

Table 2: P-values of Wilcoxon's test for all features between VOI-L from Departments A and B and between VOI-L from Departments A and A-S, before and after ComBat harmonization. Bold values are less than 0.05.

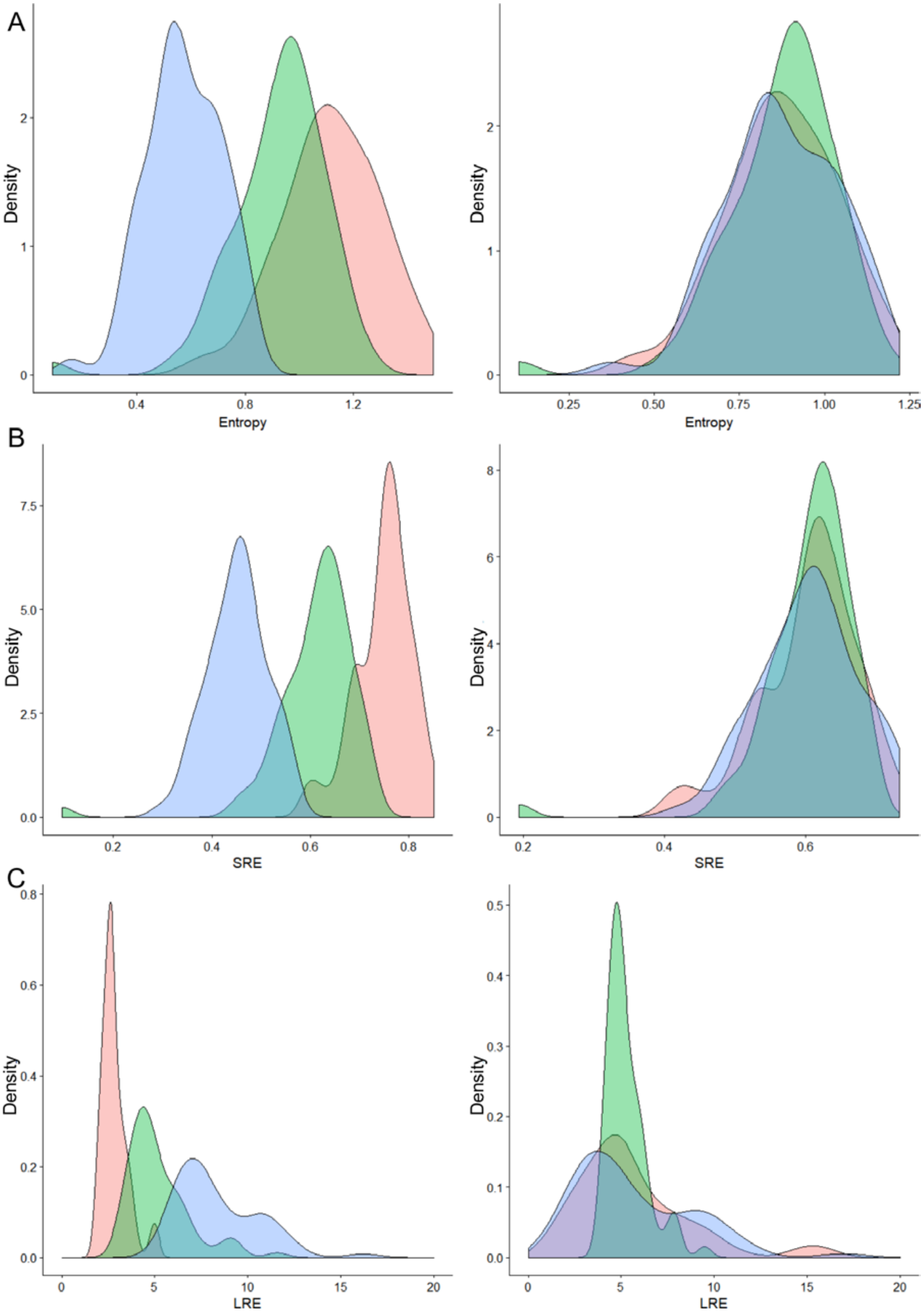
	Before ComBat						**After ComBat**			
	TN(A) vs non-TN(A)	TN(B) vs non-TN(B)	TN(A) vs TN(B)	non-TN(A) vs non-TN(B)	TN(A+B) vs non-TN(A+B)	TN(B) vs non-TN(A)	TN(A) vs TN(B)	non-TN(A) vs non-TN(B)	TN(A+B) vs non-TN(A+B)	TN(B) vs non-TN(A)
Homogeneity	0.0810	**0.0078**	0.4232	**0.0074**	**0.0014**	0.4635	0.5986	0.8737	**0.0015**	**0.0093**
Entropy	**0.0205**	**0.0410**	0.5196	0.3906	**0.0031**	0.0875	0.7405	0.9139	**0.0027**	**0.0254**
SRE	0.2175	**0.0091**	0.2995	**0.00044**	**0.0063**	0.9481	0.1294	0.8338	**0.0062**	**0.0061**
LRE	0.2618	**0.0072**	0.2814	**0.0004**	**0.0072**	0.9352	**0.0055**	0.3871	**0.0162**	**0.0004**
LGZE	**0.0005**	**0.0119**	**0.0405**	**0.0244**	**5.69e-05**	0.3786	0.1102	0.3059	**0.0002**	**0.0003**
HGZE	**0.0002**	**0.0119**	**0.0494**	**0.0282**	**3.20e-05**	0.2886	0.2814	0.3337	**2.27e-05**	**0.0058**
SUVmax	**0.0006**	**0.0111**	0.0544	**0.0278**	**7.54e-05**	0.4058	0.5717	0.7943	**4.47e-05**	**0.0072**
SUVmean	**0.0003**	**0.0139**	**0.0448**	**0.0359**	**3.20e-05**	0.2394	0.4463	0.7747	**3.05e-05**	**0.0052**
SUVpeak	**0.0004**	**0.0167**	**0.0267**	**0.0306**	**9.75e-05**	0.4736	0.3581	0.7894	**4.99e-05**	**0.0061**

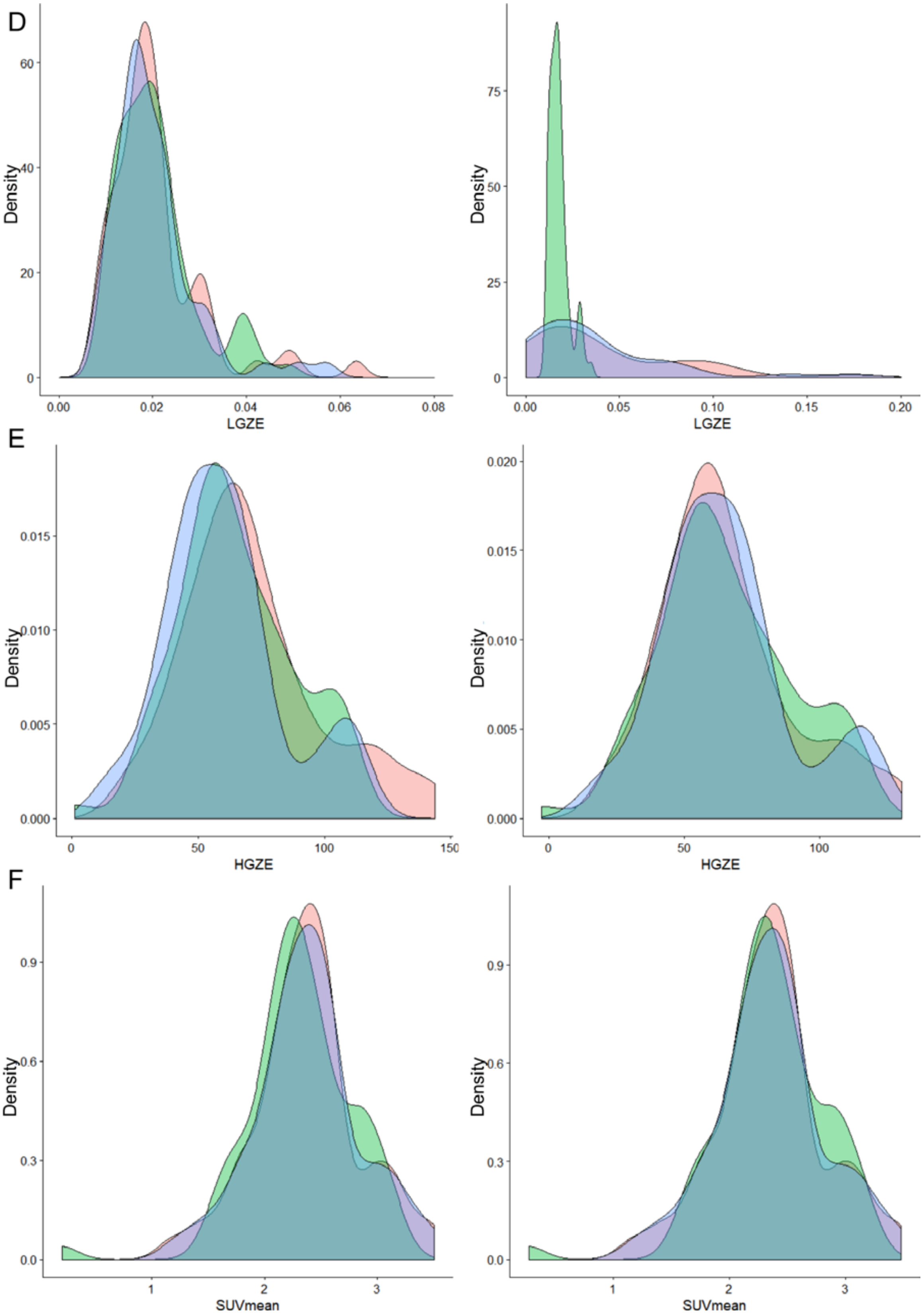
Table 3: P-values of Wilcoxon's test for all features between TN and non-TN lesions from Departments A and B, before and after ComBat harmonization. Bold values are less than 0.05.

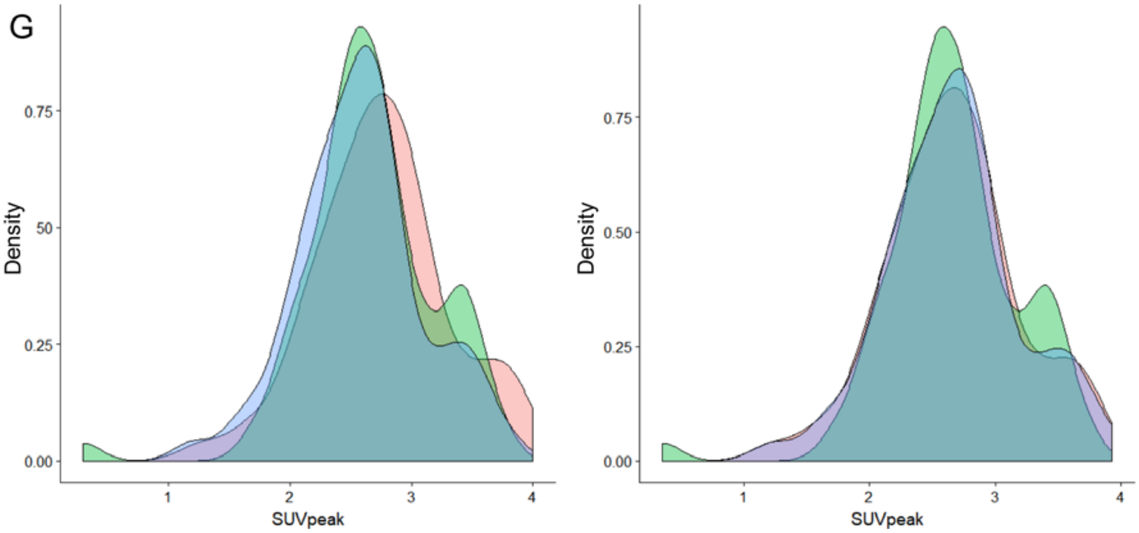
	Before ComBat	After ComBat	Thres. (B)
	Thres. (A)	Thres. (A)	
Homogeneity	0.23	0.28	0.36
Entropy	0.21	0.20	0.39
SRE	0.12	0.35	0.38
LRE	0.08	0.28	0.41
LGZE	0.07	0.33	0.39
HGZE	0.16	0.21	0.39
SUVmean	0.15	0.30	0.37
SUVmax	0.05	0.25	0.32
SUVpeak	0.05	0.36	0.37

Table 4: Youden indices for the distinction between TN and non-TN lesions from Department B when the threshold is defined based on data from Department A and when the threshold is defined based on data from Department B.

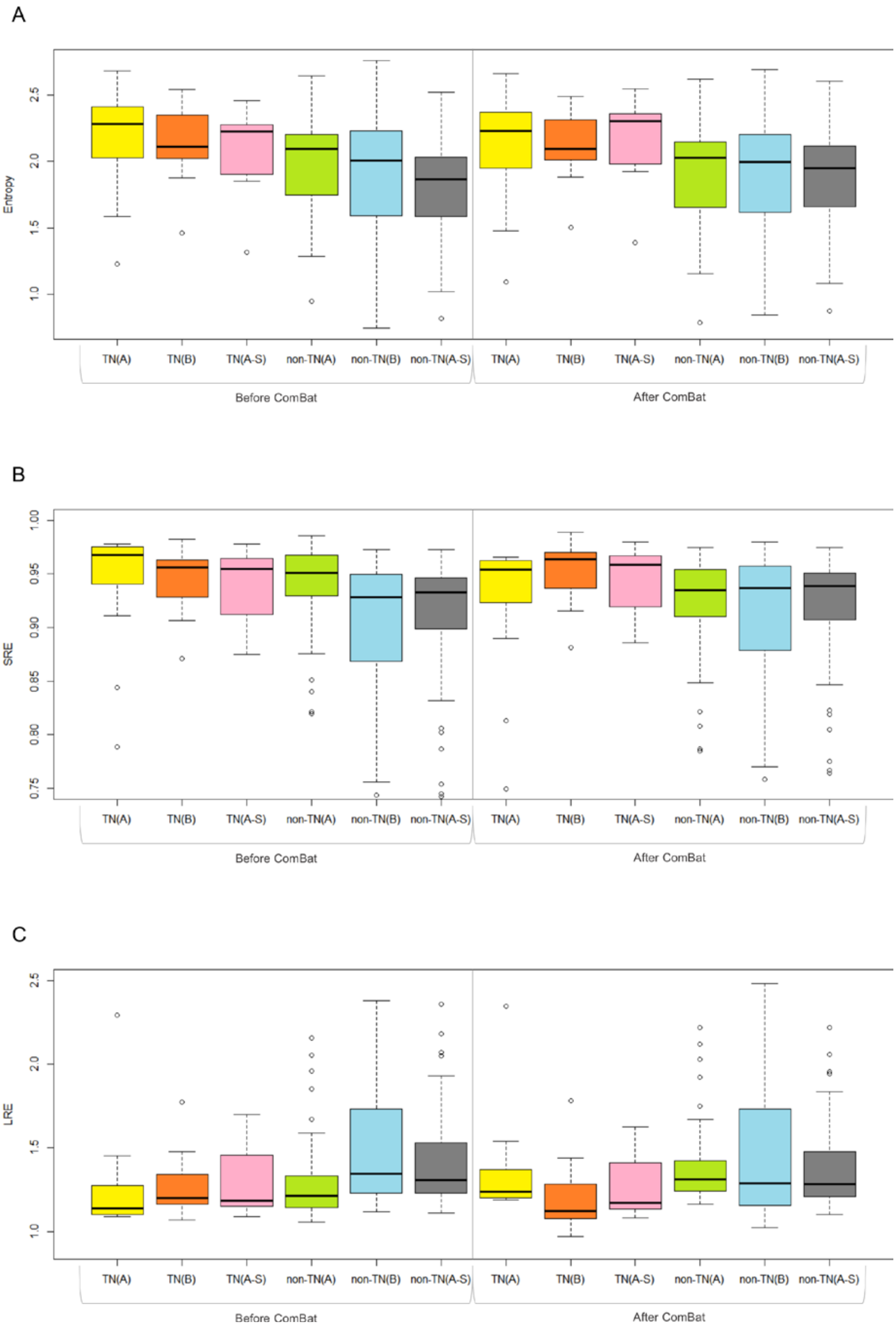
SUPPLEMENTAL FIGURE LEGEND

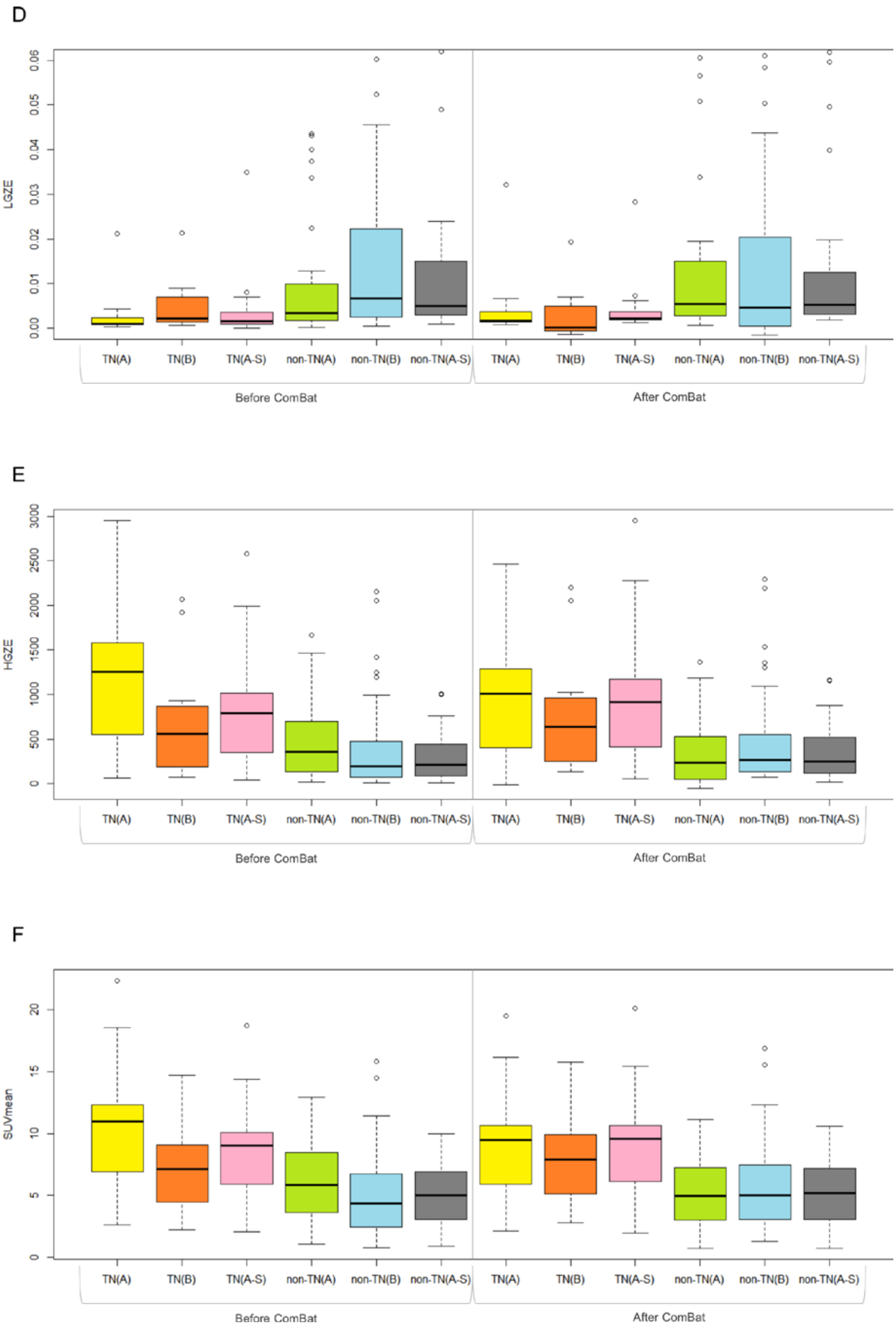



D
Density
0
20
40
60
0.00
0.02
0.04
0.06
0.08
LGZE
Density
0
25
50
75
0.00
0.05
0.10
0.15
0.20
LGZE
E
Density
0.000
0.005
0.010
0.015
0
50
100
150
HGZE
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
HGZE
F
Density
0.0
0.3
0.6
0.9
1
2
3
SUVmean
Density
0.0
0.3
0.6
0.9
1
2
3
SUVmean

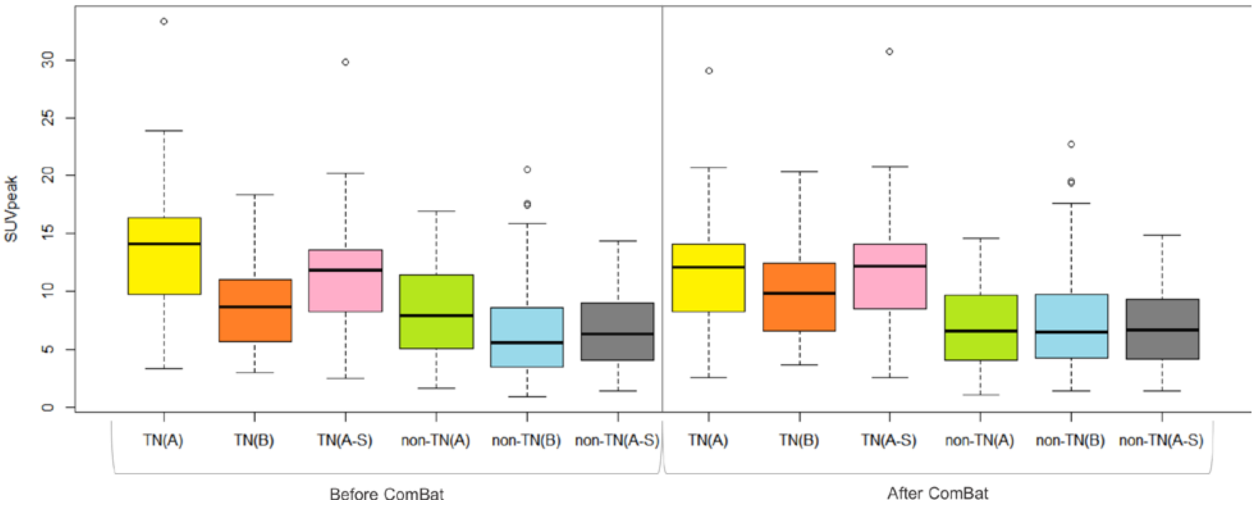


Supplemental Figure 1: Probability density function (%) of features from the liver tissue in Department A (pink), Department B (green) and Department A-S (blue), before (left) and after ComBat harmonization (right).

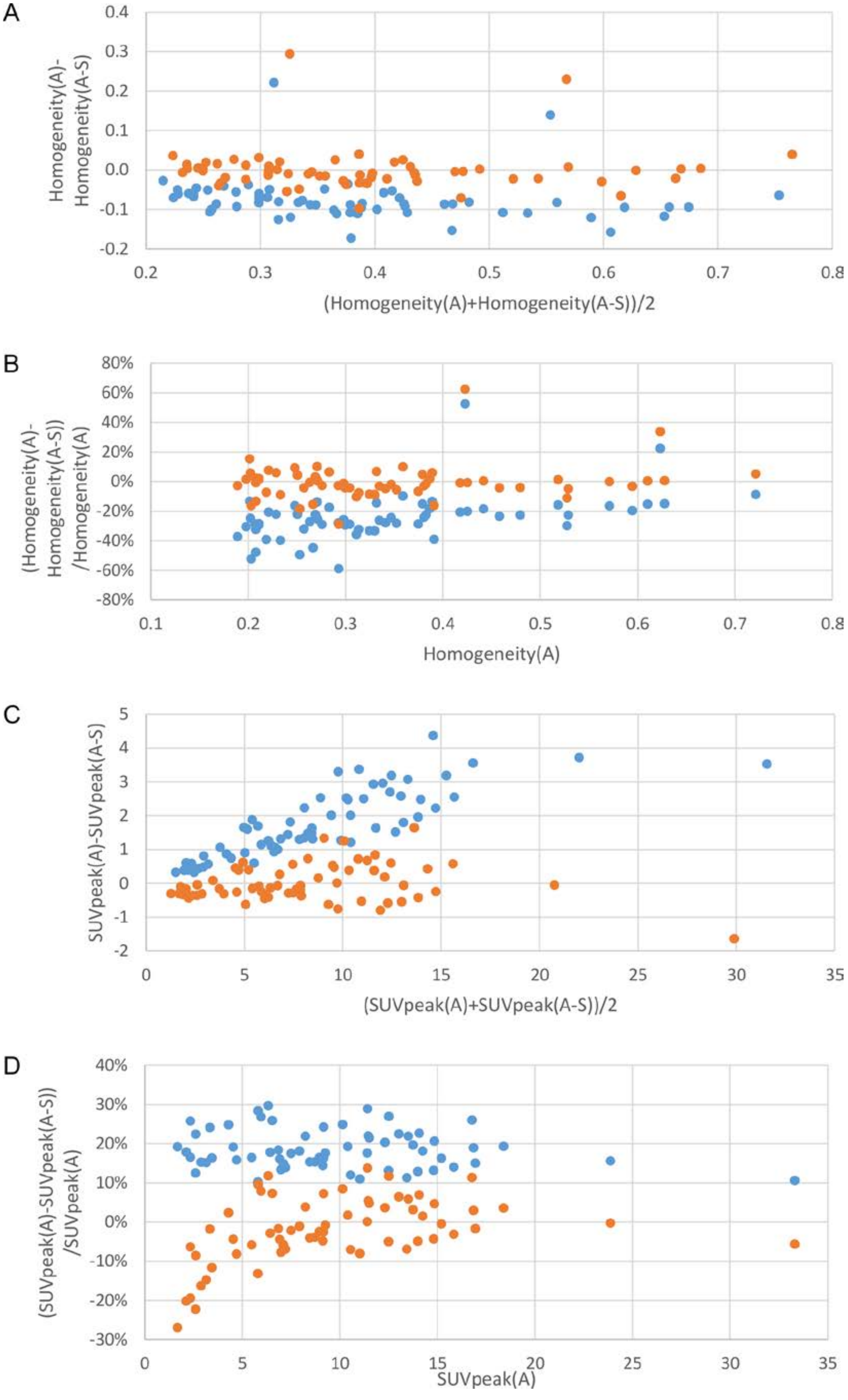

A
Entropy
1.0
1.5
2.0
2.5
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
Before ComBat
After ComBat
B
SRE
0.75
0.80
0.85
0.90
0.95
1.00
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
Before ComBat
After ComBat
C
LRE
1.0
1.5
2.0
2.5
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
Before ComBat
After ComBat


D
LGZE
0.00
0.01
0.02
0.03
0.04
0.05
0.06
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
Before ComBat
After ComBat
E
HGZE
0
500
1000
1500
2000
2500
3000
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
Before ComBat
After ComBat
F
SUVmean
0
5
10
15
20
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
TN(A)
TN(B)
TN(A-S)
non-TN(A)
non-TN(B)
non-TN(A-S)
Before ComBat
After ComBat

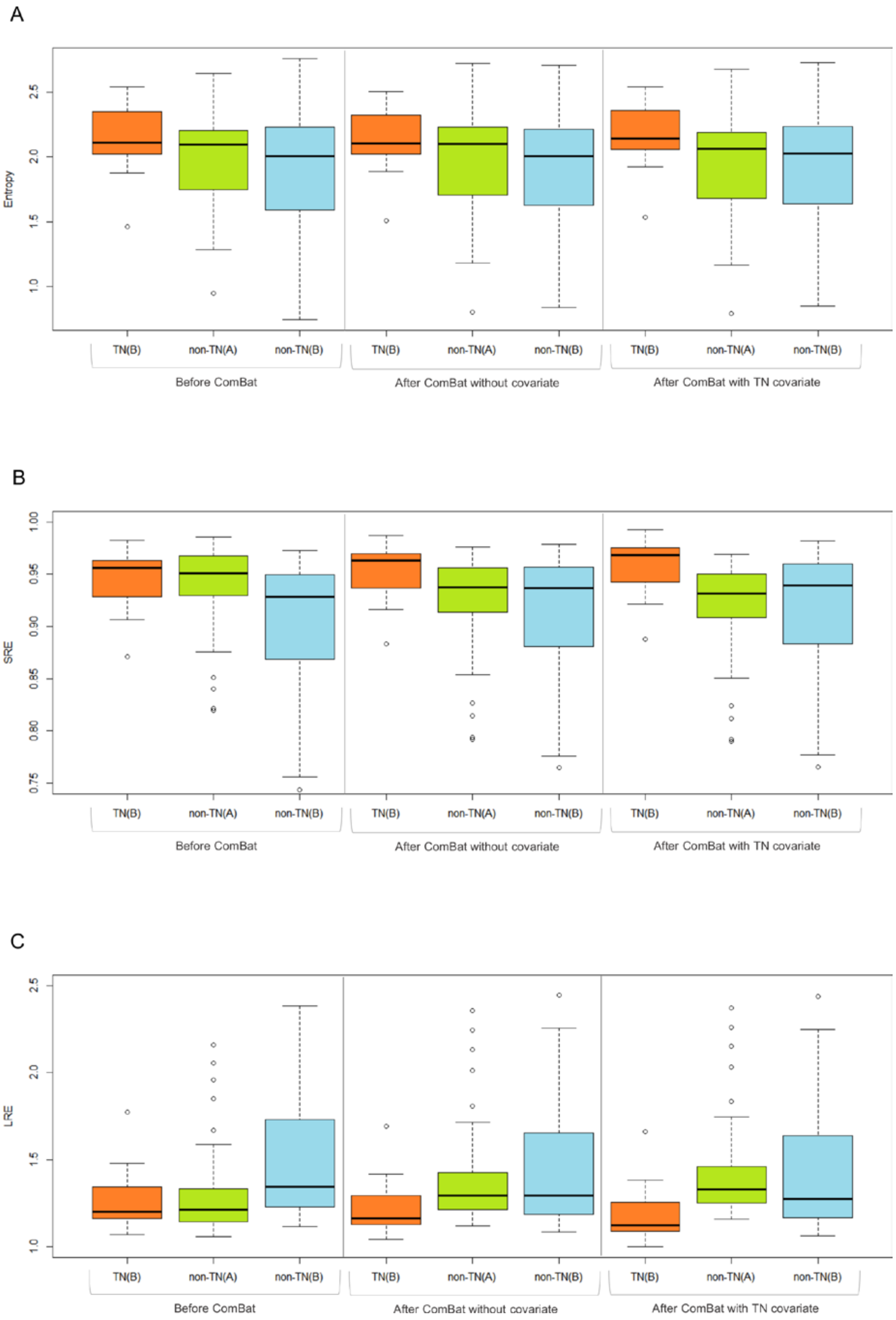
G

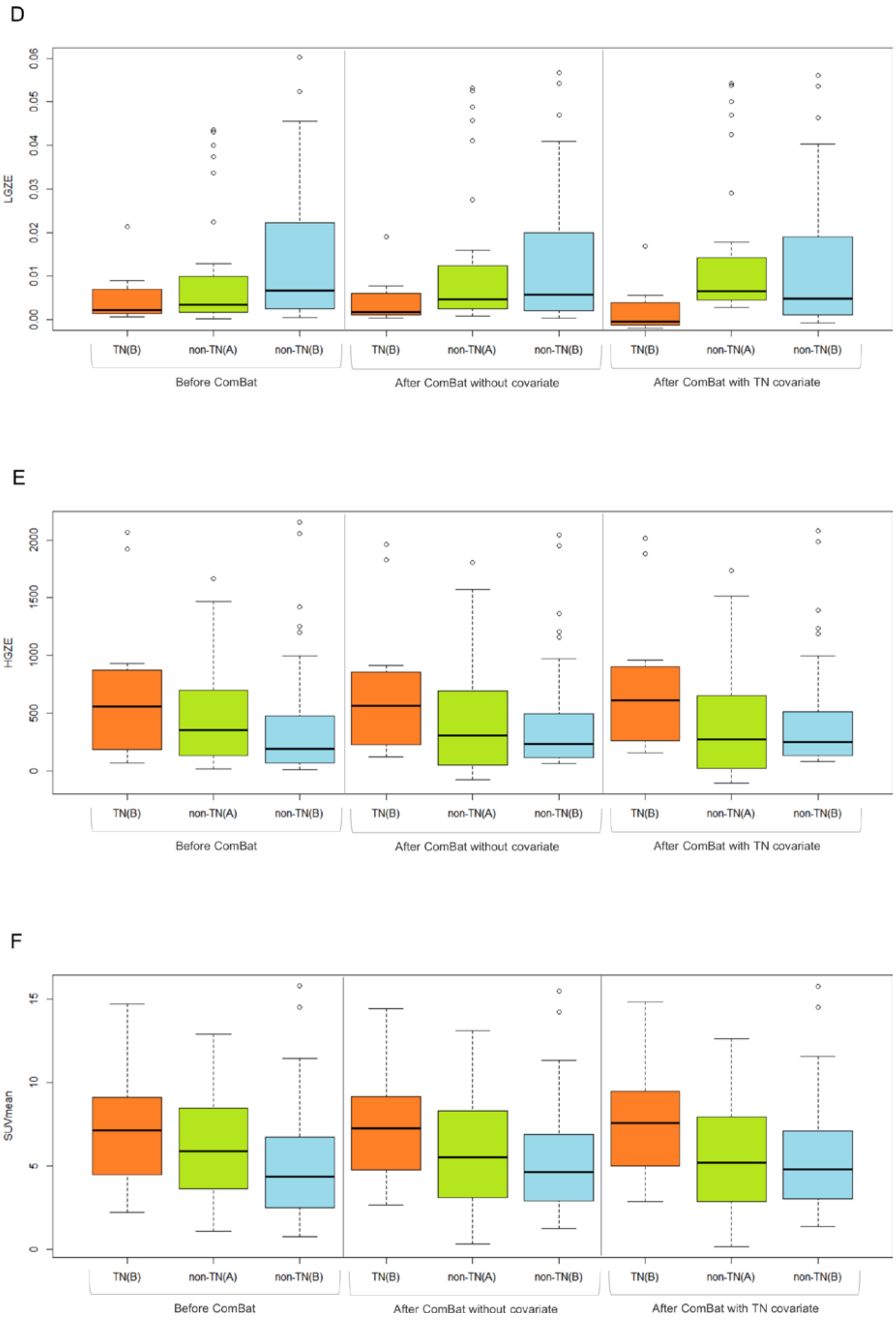


Supplemental Figure 2: Boxplots of features for triple-negative (TN) and non-triple-negative (non-TN) lesions before and after ComBat harmonization, for the three departments separately.

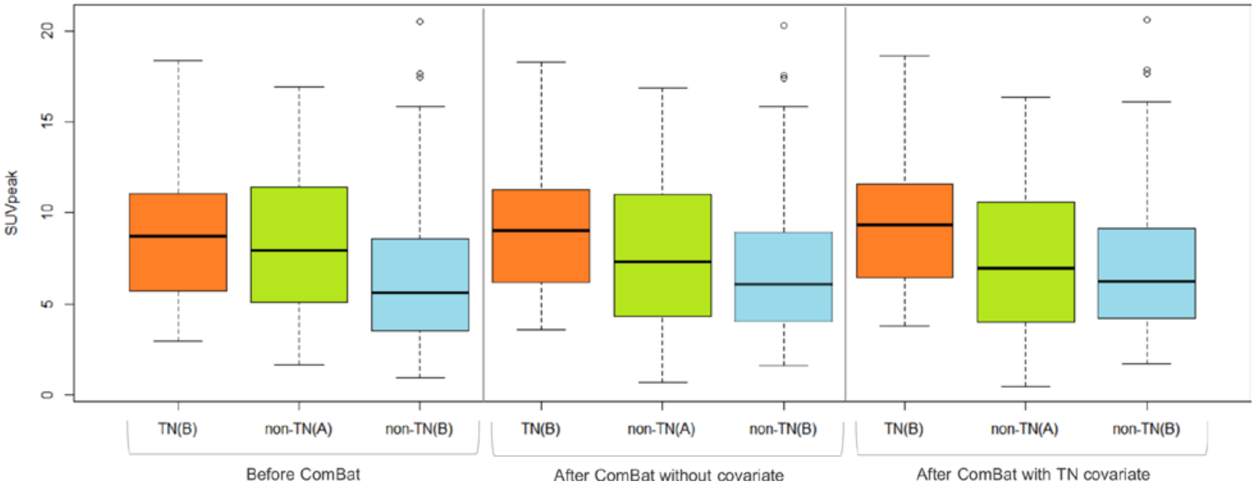

A
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
Homogeneity(A)-
Homogeneity(A-S)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
(Homogeneity(A)+Homogeneity(A-S))/2
B
80%
60%
40%
20%
0%
-20%
-40%
-60%
-80%
(Homogeneity(A)-
Homogeneity(A-S))
/Homogeneity(A)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Homogeneity(A)
C
5
4
3
2
1
0
-1
-2
SUVpeak(A)-SUVpeak(A-S)
0
5
10
15
20
25
30
35
(SUVpeak(A)+SUVpeak(A-S))/2
D
40%
30%
20%
10%
0%
-10%
-20%
-30%
(SUVpeak(A)-SUVpeak(A-S))
/SUVpeak(A)
0
5
10
15
20
25
30
35
SUVpeak(A)

Supplemental Figure 3: Bland-Altman plots (A,C) and plots of the percentage difference between features from Department A and Department A-S (B,D) before (blue) and after ComBat harmonization (orange) in tumors for Homogeneity (A,B) and SUVpeak (C,D).


A
Entropy
1.0
1.5
2.0
2.5
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
Before ComBat
After ComBat without covariate
After ComBat with TN covariate
B
SRE
0.75
0.80
0.85
0.90
0.95
1.00
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
Before ComBat
After ComBat without covariate
After ComBat with TN covariate
C
LRE
1.0
1.5
2.0
2.5
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
Before ComBat
After ComBat without covariate
After ComBat with TN covariate


D
LGZE
0.00
0.01
0.02
0.03
0.04
0.05
0.06
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
Before ComBat
After ComBat without covariate
After ComBat with TN covariate
E
HGZE
0
500
1000
1500
2000
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
Before ComBat
After ComBat without covariate
After ComBat with TN covariate
F
SUVmean
0
5
10
15
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
TN(B)
non-TN(A)
non-TN(B)
Before ComBat
After ComBat without covariate
After ComBat with TN covariate

G



Supplemental Figure 4: Boxplots of features for triple-negative (TN) and non-triple-negative (non-TN) lesions before and after ComBat harmonization without and with TN status as a covariate, for Departments A and B separately when removing all TN in Department A to determine the transformations.

	Before ComBat						**After ComBat**			
	TN(A) vs non-TN(A)	TN(A-S) vs non-TN(A-S)	TN(A) vs TN(A-S)	non-TN(A) vs non-TN(A-S)	TN(A+A-S) vs non-TN(A+A-S)	TN(A-S) vs non-TN(A)	TN(A) vs TN(A-S)	non-TN(A) vs non-TN(A-S)	TN(A+A-S) vs non-TN(A+A-S)	TN(A-S) vs non-TN(A)
Homogeneity	0.0810	**0.0103**	0.1596	**0.0006**	**0.0035**	0.7965	0.6420	0.6034	**0.0022**	**0.0124**
Entropy	**0.0205**	**0.0007**	0.3809	**0.0147**	**0.0001**	0.0587	0.5147	0.7179	**7.87e-05**	**0.0016**
SRE	0.2175	0.0525	0.1835	**0.0015**	**0.0282**	0.7724	0.4016	0.7749	**0.0241**	0.0609
LRE	0.2618	0.0867	0.1713	**0.0031**	**0.0428**	0.8329	0.0938	0.5124	**0.0457**	**0.0172**
LGZE	**0.0005**	**0.0003**	0.1381	**0.0472**	**1.77e-06**	**0.0327**	0.2099	0.6899	**1.50e-06**	**0.0041**
HGZE	**0.0002**	**0.0002**	0.0938	**0.0284**	**1.01e-06**	**0.0289**	0.7804	0.6246	**6.37e-07**	**0.0004**
SUVmax	**0.0006**	**0.0002**	0.0731	**0.0050**	**5.13e-06**	0.1737	0.9556	0.9820	**1.10e-06**	**0.0005**
SUVmean	**0.0003**	**0.0002**	0.1381	0.0820	**9.52e-07**	**0.0150**	0.8965	0.9880	**4.49e-07**	**0.0002**
SUVpeak	**0.0004**	**0.0002**	0.2703	0.0660	**1.23e-06**	**0.0196**	0.8672	0.9940	**6.01e-07**	**0.0003**

Supplemental Table 1: P-values of Wilcoxon's test for all features between TN and non-TN lesions from Departments A and A-S, before and after ComBat harmonization. Bold values are less than 0.05.

	Before ComBat	**After ComBat without TN Cov**	**After ComBat with TN Cov**
Homogeneity	0.4635	**0.0265**	**0.0039**
Entropy	0.0875	0.1232	**0.0254**
SRE	0.9410	**0.0203**	**0.0002**
LRE	0.9352	**0.0039**	**7.64e-05**
LGZE	0.3786	**0.0132**	**1.80e-06**
HGZE	0.2886	0.1041	**0.0254**
SUVmax	0.4058	0.1192	**0.0423**
SUVmean	0.2394	0.0939	**0.0359**
SUVpeak	0.4736	0.1403	**0.0478**

Supplemental Table 2: P-values of Wilcoxon's tests for all features between TN in Department B and non-TN in Department A when the proportion of TN tumors was very different in the two departments, before and after ComBat harmonization without and with TN as a covariate of interest. Bold values are less than 0.05.