Check for updates

Human Brain Mapping

WILEY

RESEARCH ARTICLE OPEN ACCESS

Assessment of ComBat Harmonization Performance on Structural Magnetic Resonance Imaging Measurements

Emma Tassi[1,2] | Anna Maria Bianchi[2] | Federico Calesella[3,4] | Benedetta Vai[3,4] | Marcella Bellani[5] | Igor Nenadić[6] | Fabrizio Piras[7] | Francesco Benedetti[3,4] | Paolo Brambilla[1,8] | Eleonora Maggioni[1,2]

[1]Department of Neurosciences and Mental Health, Fondazione IRCS Cà Granda Ospedale Policlinico, Milano, Italy | [2]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy | [3]Unit of Psychiatry and Clinical Psychobiology, Division of Neuroscience, IRCCS Ospedale San Raffaele, Milano, Italy | [4]University Vita-Salute San Raffaele, Milano, Italy | [5]Section of Psychiatry, Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy | [6]Department of Psychiatry and Psychotherapy, Philipps-University Marburg/ Marburg University Hospital, Marburg, Germany | [7]Fondazione Santa Lucia, Roma, Italy | [8]Department of Pathophysiology and Transplantation, University of Milan, Milano, Italy

Correspondence: Paolo Brambilla (paolo.brambilla1@unimi.it)

Received: 10 May 2024 | **Revised:** 16 September 2024 | **Accepted:** 15 November 2024

Funding: This work was supported by the Moodlearning grant from Italian Ministry of Health, GR-2018-12367789. P.B. was partially supported by grants from the Italian Ministry of Education and Research - MUR ('Dipartimenti di Eccellenza' Program 2023–27 - Dept. of Pathophysiology and Transplantation, Università degli Studi di Milano), the Italian Ministry of Health (Hub Life Science- Diagnostica Avanzata, HLS-DA, PNC-E3-2022-23683266– CUP: C43C22001630001 / MI-0117; Ricerca Corrente 2024), by the Fondazione Cariplo (Made In Family, grant number 2019–3416), and by the ERANET Neuron JTC 2023 (ERP-2023-23684211 - ERP-2023-Neuron-ResilNet). A.M.B received funding from the NextGenerationEU–National Recovery and Resilience Plan (NRRP), Mission 4, "Education and Research"—Component 2, "From research to business," Investiment 3.1—Call for tender 3264 (December 28, 2021) of the Italian Ministry of University and Research, under the project titled "EBRAINS-Italy (European Brain ReseArchINfrastructureS-Italy)" (Project code: IR0000011, Concession Decree 117 of June 21, 2022). E.M. was partly supported by the Italian Ministry of University and Research (PRIN 2022, 2022RXM3H7).

Keywords: ComBat | cortical thickness | gray-matter volume | harmonization | MRI

ABSTRACT

Data aggregation across multiple research centers is gaining importance in the context of MRI research, driving diverse high-dimensional datasets to form large-scale heterogeneous sample, increasing statistical power and relevance of machine learning and deep learning algorithm. Site-related effects have been demonstrated to introduce bias in MRI features and confound subsequent analyses. Although Combating Batch (ComBat) technique has been recently reported to successfully harmonize multi-scale neuroimaging features, its performance assessments are still limited and largely based on qualitative visualizations and statistical analyses. In this study, we stand out by using a robust cross-validation approach to assess ComBat performances applied on volume- and surface-based measures acquired across three sites. A machine learning approach based on Multi-Class Gaussian Process Classifier was applied to predict imaging site based on raw and harmonized brain features, providing quantitative insights into ComBat effectiveness, and verifying the association between biological covariates and harmonized brain features. Our findings showed differences in terms of ComBat performances across measures of regional brain morphology, demonstrating tissue specific site effect modeling. ComBat adjustment of site effects also varied across regional level of each specific volume-based and surface-based measures. ComBat effectively eliminates unwanted data site-related variability, by maintaining or even enhancing data association with biological factors. Of note, ComBat has demonstrated flexibility and robustness of application on unseen independent gray matter volume data from the same sites.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Human Brain Mapping* published by Wiley Periodicals LLC.

1 | Introduction

Multimodal neuroimaging data sharing across research communities and sites has become increasingly common, creating a growing trend toward forming large-scale datasets from heterogeneous samples. Moreover, the integration of data across multiple sites represents one of the increasingly exploited strategies for application of artificial intelligence techniques, since an increase of numerosity of the training set can help reaching better accuracies in the learning process. However, multisite neuroimaging data are biased by undesirable non-biological sources of variability, attribute to the use of different imaging scanners and parameters, which can affect reproducibility and consistency of subsequent analysis, possibly leading to erroneous findings like misclassifications. Since the magnitude of site's effects can be larger than diagnosis effects or other effects of interest, adequately addressing these confounding factors becomes of paramount importance. A recent software tool that is increasingly used to deal with undesired data heterogeneity is Combating Batch (ComBat) effects (Johnson, Li, and Rabinovic 2007; Pomponio et al. 2020) originally applied for removing "batch effects" from genomic data and then adapted for neuroimaging data. ComBat is now commonly used as preprocessing step to remove site effects while preserving site-specific biological variability. Specifically, Combat harmonization applies a location and scale adjustment model considering input raw data as a linear combination of biological and non-biological terms, accounting for additive and multiplicative site contributions. Recent neuroimaging researches reported ComBat to successfully harmonize multi-scale features from different neuroimaging data types, for example, diffusion tensor imaging (Fortin et al. 2017), functional magnetic resonance imaging (fMRI) (Yu et al. 2018; Tassi et al. 2023), and structural MRI (sMRI) (Pomponio et al. 2020; Fortin et al. 2018; Beer et al. 2020; Radua et al. 2020). As regards the latter technique, ComBat performances have been tested on regional cortical thickness (CT; Fortin et al. 2018; Beer et al. 2020; Radua et al. 2020), regional gray matter volume (GMV; Pomponio et al. 2020; Radua et al. 2020), voxel-based morphometry (Takao, Hayashi, and Ohtomo 2011), and regional surface area (Radua et al. 2020) features. Nevertheless, quantitative evidence for ComBat effectiveness in removing site-related variability while preserving the effects of interest is still limited, with some studies showing no added values of ComBat over standard linear regression approaches (Zavaliangos-Petropulu 2019). Moreover, within these studies, the assessment of ComBat performances was based on statistical analyses on site differences in the analyzed features, or by exploratory visualization of feature-site associations aimed to compare how much raw and harmonized data are clustered by imaging sites (Yu et al. 2018; Fortin et al. 2018). In addition, clustering was employed to discover any site-related clusters in raw and harmonized feature sets (Yamashita et al. 2019). So far, only one study applied supervised machine learning (ML) techniques to evaluate ComBat performances (Fortin et al. 2018). Furthermore, a consensus on ComBat pipeline has not been reached, since some studies estimate and remove site effects in the same dataset (Pomponio et al. 2020; Yu et al. 2018; Fortin et al. 2018; Beer et al. 2020; Yamashita et al. 2019) other employ cross-validation (CV) frameworks that separate fitting and application of harmonization parameters in different feature sets (Radua et al. 2020). Specifically, a recent ENIGMA study (Radua et al. 2020) introduced the idea of adapting ComBat functions allowing harmonization of a test set based on parameters estimated in a training set, thus addressing the inclusion of ComBat within a CV approach.

By considering previous studies evaluating reliability of ComBat technique (Fortin et al. 2017; Fortin et al. 2018), with this multisite study, we aimed to quantitatively and comprehensively address ComBat performances on brain morphological measures using both volume-based and surface-based measures and for each of them diverse atlases via a robust CV approach. To this end, we applied a ML approach based on Multi-Class Gaussian Process Classifier (MCGPC) aimed to predict imaging site from the raw and harmonized brain features, revealing quantitative insights into site effect removal after ComBat application. Further, we estimated, before and after harmonization, the association between biological covariates and brain features using linear regression analyses as well as tailored ML, providing information on ComBat capability to effectively control for biological feature variability. Notably, by the adoption of a CV framework, we tested ComBat performances in harmonizing unseen independent brain features from the same centers across surface-based and volume-based measures of brain morphology. This flexibility and robustness analysis of ComBat on independent test sets would support the emerged importance in clinical practice of extracting deviations created by differences in acquisition site across multisite clinical datasets and apply them on independent datasets from the same site, without replicating site effects estimation with an independent algorithm. Thus, a detailed evaluation of ComBat performances within a CV framework would be considered as an important adjunctive step for facilitate screening deviations effect in unseen independent dataset from the same site.

We hypothesized that (1) ComBat can be used to remove unwanted site-related variability in the data while preserving the site-unrelated biological variability; (2) after training on a multisite dataset, ComBat can be used to harmonize new independent data from the same sites by applying the coefficients estimated on the training set; and (3) ComBat estimates the highest coefficients for the site associated with the maximum distance from the others in terms of feature values before harmonization. Aside verifying foundational capability of ComBat, this study aims to highlight the importance of evaluating ComBat performances by applying a tissue-specific analysis and employing a variety of metrics. Further, beyond the functionality of ComBat on harmonizing independent test sets, we aim to perform a detailed and comparative analysis evaluating the flexibility and robustness of ComBat performances within a CV framework across different regional-based measures.

2 | Material and Methods

2.1 | Dataset

The dataset included in our analysis is composed of sMRI data acquired from 294 healthy volunteers (154 females, 140 males, 32.98 ± 8.71 years) across four sites of the StratiBip network, which stemmed from the ENPACT network (Delvecchio et al. 2021; Maggioni et al. 2017). The four sites are: Azienda Ospedaliera Universitaria Integrata of Verona, Italy (AOUV,

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Site 1), University Hospital of Jena, Germany (JUH, Site 2), IRCCS Ospedale San Raffaele of Milan, Italy (OSR, Site 3) and Fondazione IRCCS Santa Lucia of Rome, Italy (FSL, Site 4). The sample characteristics are reported in Table 1. Structural brain images were acquired using 3 T MRI scanners and T1-weighted sequences that differed across the four sites, described in Table 2. Using a holdout pipeline, the dataset was split into training (75%) and test (25%) sets by maintaining the balance across sites, obtaining 220 subjects in training set and 74 subjects in test set.

For continuous variables, mean standard deviations are indicated. Kruskal Wallis test (KW) for age differences across sites: $p = 0.39$; Chi-squared test for sex differences: $p = 0.83$. Kruskal Wallis test for inter-site differences of age for both training and test: (i) training: $p = 0.89$, (ii) test: $p = 0.17$. Fisher's test for inter-site differences of sex for training and test: (i) training: $p = 0.28$, (ii) test: $p = 0.38$; MPRAGE: T1 Magnetization Prepared Rapid Gradient Echo; FFE: T1-Fast Field Echo.

2.2 | sMRI Data Preprocessing and Feature Extraction

The multisite T1-weighted images were processed on the same workstation using MATLAB R2021b (The MathWorks Inc., Natick) for the extraction of volume- and surface-based brain morphology measurements. The computational anatomy toolbox (CAT12: https://neuro-jena.github.io/cat/) within the Statistical Parametrical Mapping software (SPM12: http://www.fil.ion.ucl.ac.uk/spm/) (Penny et al. 2011) was used for image preprocessing and estimation of regional GMV and CT. The preprocessing included tissue segmentation, spatial normalization to the standard Montreal Neurological Institute space template, brain tissue segmentation, adjustment of segmented tissues for volume alterations during registration (i.e., modulation), and spatial smoothing via convolution with a 3D Gaussian kernel (6 mm). Thus, smoothed, modulated, and normalized GM volumes were employed for the extraction of GMV and CT from brain regions defined according to probabilistic atlases. For each subject, mean GMV values were extracted for regions of the volume-based Cobra ($n = 52$) and Neuromorphometrics ($n = 136$) atlases (Park 2014), whereas mean CT values were extracted for regions of the surface-based Desikan-Killiany ($n = 72$) and Destrieux ($n = 152$) atlases (Destrieux et al. 2010; Desikan 2006). Among them, GMV features from 42 regions of interest (ROIs) of Cobra atlas and 122 ROIs of Neuromorphometrics atlas were extracted, after having excluded the ROIs with only white matter (WM). CT features from 144 ROIs of Destrieux atlas and 64 ROIs of Desikan-Killiany atlas were selected based on the availability of CT regional measures across subjects. In addition, global measures related to cerebrospinal fluid, gray matter, and WM volumes and total intracranial volumes (TIVs) were extracted.

2.3 | ComBat-Based Feature Harmonization

The removal of site effects from GMV and CT features was performed via the ComBat harmonization procedure proposed by Johnson, Li, and Rabinovic (2007). This technique considers the data as a linear combination of site effects, composed of multiplicative (δ) and additive (γ) effects, and biological covariates. ComBat harmonization procedure is as follows. First, each feature is standardized to obtain similar overall mean and variance across batches (i.e., site). The standardized data were assumed to be normally distributed, and the mean and variance of site-specific distributions were identified as site effects estimators, following Normal and Inverse Gamma prior distributions, respectively. Hyperparameters of such prior distributions, were calculated empirically from the previously standardized data with the use the methods of moments. As last, non-harmonized standardized features were adjusted based on the site effects and resulting in harmonized feature values.

The model can be written as follows:

$$y_{ijv} = \alpha_v + X_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv} \epsilon_{ijv} \tag{1}$$

where y_{ijv} is the feature value from the regional structural atlas (i.e., GMV, CT) of the imaging site i, participant j and feature values v. α_v is the average feature CT or GMV for the reference site i for feature v (relative to a single ROI), X_{ij}^T is the design matrix for covariates of interest, and β_v is the vector of regression coefficient related to X. The error terms ϵ_{ijv} follow a normal distribution with mean 0 and variance σ_v^2. γ_{iv}^* and δ_{iv}^* indicate additive and multiplicative site effects of imaging site i for feature value v. The ComBat-harmonized values can be defined as:

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \alpha_v + X_{ij}^T \beta_v - \gamma_{iv}^*}{\delta_{iv}^*} + \alpha_v + X_{ij}^T \beta_v \tag{2}$$

In our study, ComBat harmonization was performed using a publicly available MATLAB-based package (https://github.com/Jfortin1/ComBatHarmonization). Before the ROI-based feature harmonization, the TIV was harmonized with the other global volumes, while considering age and sex as biological covariates. ComBat was then applied to GMV and CT features, considering the features from each atlas as a separate feature set, within a CV framework. ComBat was first applied to the training sets, while considering age, sex contributions for GMV and age, sex,

TABLE 1 | Demographic information of the entire sample.

ID	Location	*N* of subjects	Age range (years)	Age mean ± std. (years)	*N* of males/females
Site1	AOUV	73	25–54	31.72 ± 6.08	35/38
Site2	JUH	74	25–55	32.02 ± 8.05	34/40
Site3	OSR	67	19–62	35.03 ± 12.74	35/32
Site4	FSL	80	24–47	33.29 ± 6.82	36/44

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and harmonized TIV contributions for CT. By considering the degree of nuisance correlation of CT with TIV, TIV was included as covariate for ComBat harmonization applied on CT.

At this stage, additive and multiplicative site effects were estimated. The GMV and CT test sets were then harmonized by applying the ComBat site-effect correction coefficients estimated on the relative training set. A flowchart indicative of our ComBat harmonization pipeline is reported in Figure 1.

2.4 | Sites Effect

2.4.1 | Visual Inspection of ComBat Coefficients

An exploratory visualization analysis was employed to evaluate site effects in terms of multiplicative and additive coefficients. In addition, to quantify site effects, for each site i the mean values of δ and γ coefficients across features of each type were extracted (i.e., δ_{mean}, γ_{mean}) as follows:

TABLE 2 | Description of MRI scanner and sequence parameters.

Site	Location	Manufacturer	Sequence	Magnetic field (T)	Matrix size	Voxel size (mm×mm×mm)
1	AOUV	Magnetom Allegra Syngo (Siemens, Erlangen, Germany)	T1 MPRAGE	3	256×256×160	1.00×1.00. ×1.00
2	JUH	Siemens Tim Trio (Siemens, Erlangen, Germany)	T1 MPRAGE	3	256×256×192	1.00×1.00. ×1.00
3	OSR	Philips Intera (Philips, Best, the Netherlands)	FFE	3	256×256×220	0.9×0.9×0.8
4	FSL	Philips Medical Systems (Philips, Best, the Netherlands)	T1 MPRAGE	3	432×432×190	0.54×0.54×0.9

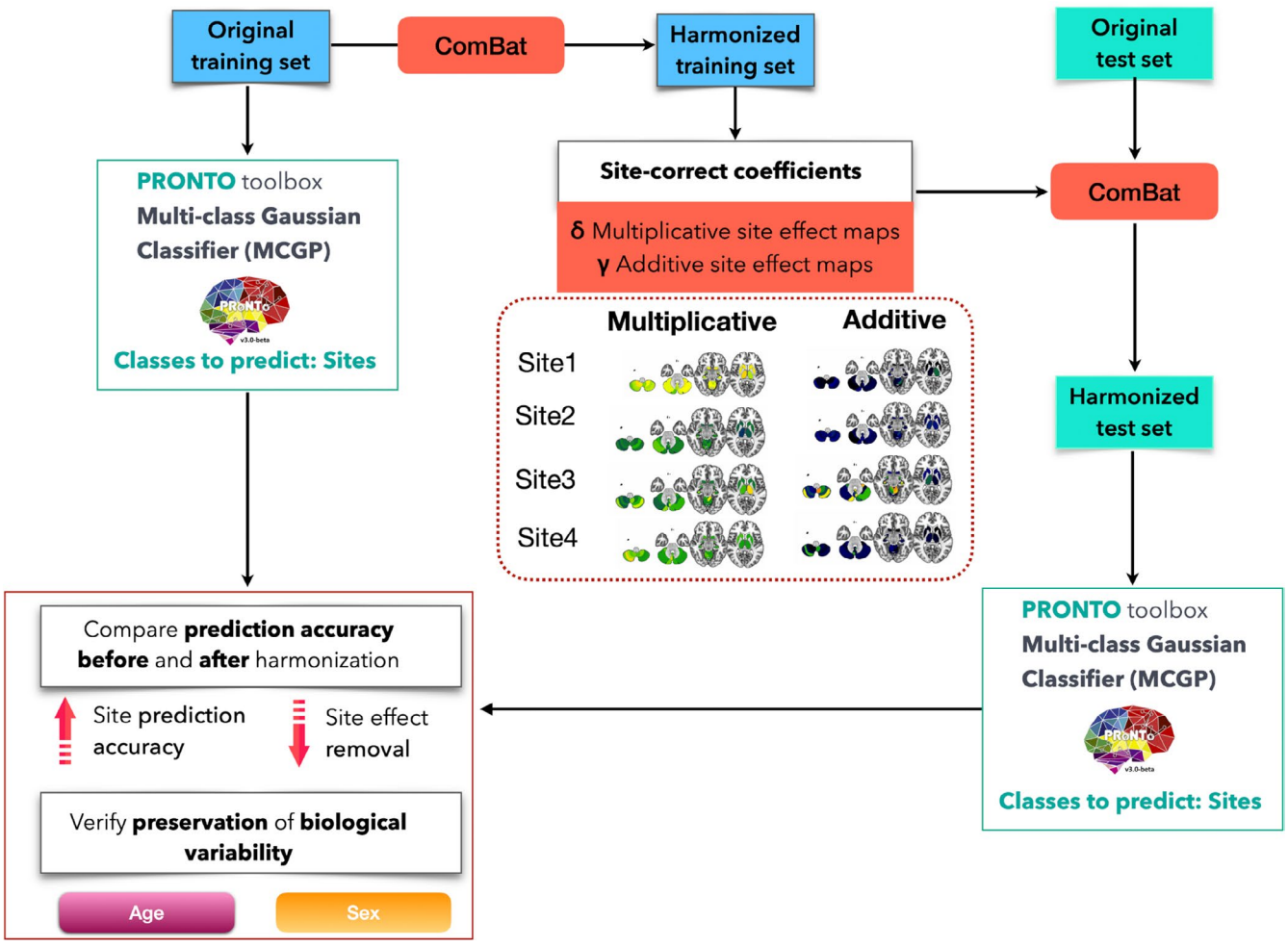

FIGURE 1 | Flowchart describing the cross-validation approach to assess ComBat performances applied on volume- and surface-based measures acquired across three sites.

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\delta_{mean,iv} = \frac{\sum_{n=1}^{N} \delta_n}{N} \quad (3)$$

$$\gamma_{mean,iv} = \frac{\sum_{n=1}^{N} \gamma_n}{N} \quad (4)$$

where N indicates the number of regions included in the ROI-based feature v. δ_{mean} *and* γ_{mean} are indicative of the total average magnitude of site effects. Thus, sites were rated based on the amounts of the estimated site effects. As last, the total average magnitude of site's estimators ($\delta_{mean} + \gamma_{mean}$) was also extracted, representing a composite metric indicative of the overall site effects, summarizing additive and multiplicative sources.

2.5 | ComBat Performance Test: Removal of Sites Effect

ComBat performances were assessed using a combination of previously proposed and novel approaches, all of which compared a specific metric before and after ComBat application. The use of multiple metrics allowed us to test ComBat in terms of site effect removal (site-related metrics) and biological variability preservation (biology-related metrics). As for ComBat application, all these tests were performed separately for each GMV (Cobra and Neuromorphometrics) and CT (Destrieux and Desikan-Killiany) atlas.

2.5.1 | Feature Comparison Among Sites

Nonparametric KW tests were applied to quantify any difference among sites in the median values of all features within each feature set, before and after applying ComBat harmonization, separately for training and test sets. As follow, if significant differences emerged ($p < 0.05$, Bonferroni corrected with $n = 4$, number of feature sets), post hoc multiple pairwise comparison results were extracted.

2.5.2 | Visual Inspection of Site Effect via Principal Component Analysis

For each set of features, a principal component analysis (PCA) was performed in the training and test sets, before and after ComBat harmonization. The scatterplot of the first two principal component (PC) scores was visualized by representing different sites with different colors to qualitatively assess the amount of site-related variation captured by the PCA. In case of significant feature differences among sites, we expected the PC scores to be associated with site, that is, showing samples from the same site clustered together in the PCA scatterplot. For each feature set, we also visually compared the scatterplot of the first PC scores before and after ComBat harmonization in the training and test sets.

2.5.3 | Site Classification Models

PRONTO (Schrouff 2013) MCGPC was adapted for application to ROI-based features by employing in-house MATLAB scripts and used to classify the site from each feature set, using separately training and test sets, before and after ComBat harmonization. The MCGP, which operates a probabilistic kernel-based algorithm, results well-suited for sorting instances into multiple categories and thus informing predictive probabilities of multiclass membership (Rasmussen and Williams 2006; Wegrzyn et al. 2015). Indeed, MCGP strength lies in its ability to handle multi-class problems and the capacity to yield probabilistic predictions.

The comparison of site classification accuracies provided a quantitative measure of the extent of site effect removal after ComBat application. Specifically, good harmonization performance should reduce the accuracy of site classification obtained from the harmonized features with respect to non-harmonized ones. For the training set, the MCGPC training classification performance was assessed through a 10-fold CV scheme. Further, MCGPC test classification performance was evaluated by a k-fold CV automatically chosen by PRONTO. Classification performances were measured using standard metrics including balanced accuracies (BAs), class- (i.e., site-) and specific accuracies (CA). Specifically, CA is computed by considering disjoint subsets of the whole testing data, where each subset contains only test samples from one class (i.e., site). Based on CA sets, BA could be computed as the average accuracy obtained on either class, thus giving indices of performance that consider the different size of the groups. A single value of BA, CA, sensitivity, specificity, and precision were extracted as the average values across folds.

The performance significance was tested by permuting the outcome label, represented by the imaging site ($N = 1000$). p-Values were validated based on permutation tests. The differences between the drops in BA from before to after ComBat harmonization obtained in training and test sets (i.e., $\Delta_{BA\ drops} = BA\ drops_{train} - BA\ drops_{test}$) were extracted to assess the differential performance of ComBat from the training to the independent test set. Furthermore, sites were ranked in terms of the average CA drop after ComBat harmonization across feature sets.

2.6 | Preservation of Biological Variability

2.6.1 | Linear Regression Analysis

ComBat harmonization on the association between biological factors with both individual and median values of the features in each feature set, and separately for training and test, was quantitatively assessed via linear regression using in-house MATLAB scripts. Before and after Combat application, linear regression models were fitted, using age and sex as independent variables, and each feature as dependent variable. Harmonized and non-harmonized feature sets were compared in terms of T-statistics and relative p-values, as well as the percentage of variation in the dependent variable (i.e., the selected feature) explained by age and sex, that is, the coefficient of determination R^2. In these analyses, significance threshold was set to $p = 0.05$. No multiple-hypothesis testing was performed, given the only interest in comparing the associations before and after ComBat harmonization, and not in testing the absolute significance of these associations.

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.6.2 | Kernel Ridge Regression for Age Prediction

We quantified the capability to predict age from harmonized and non-harmonized values of each feature set using Kernel ridge regression (KRR) algorithm from PRONTO toolbox, separately for training and test sets. Mean-squared error (MSE) was employed as performance metric and compared before and after ComBat harmonization. A fivefold CV on leave-one-subject-out with hyperparameter optimization was used for the CV internal loop, whereas leave-one-subject-out scheme was used for the CV external loop.

2.6.3 | Binary Support Vector Machine for Sex Prediction

We quantified the capability to classify sex from harmonized and non-harmonized values of each feature set using binary support vector machine (SVM), separately for training and test sets. Sex classification accuracy was compared before and after ComBat harmonization. For the CV inner loop, hyperparameter optimization was performed through fivefold CV, while 10-fold CV on leave-one-subject-out was used for outer loop to estimate model performance. A selected balanced sample of males and females was employed for both training and test application.

3 | Results

In this section, additive and multiplicative site effects estimated by ComBat are first inspected and discussed. Further, the main findings of the ComBat performance tests in terms of site effect removal and preservation of biological variability are presented.

3.1 | Sites Effect

3.1.1 | Visual Inspection of ComBat Coefficients

Figure S1 shows ComBat multiplicative (δ) and additive (γ) coefficient absolute values (magnitude) for GMV features from Neuromorphometrics atlas and CT features from Desikan-Killiany atlas.

For each feature set and site, the features associated with the highest multiplicative and additive coefficient values are listed in Tables S1–S5, together with the average magnitude γ_{mean} and δ_{mean} coefficients across features, as well as the total average magnitude of site's estimators ($\delta_{mean} + \gamma_{mean}$). For GMV features from both Neuromorphometrics and Cobra atlases, the site with the highest was Site 1, while Site 2 was associated with the lowest one. In both feature sets, the highest value was found in Site 3; conversely, the lowest were found in Site 4 and Site 1 for the Neuromorphometrics and Cobra atlases, respectively. In both feature sets, the site with the highest $\delta_{mean} + \gamma_{mean}$ was Site 3. The CT features from both Desikan-Killiany and Destrieux atlases were characterized by highest in Site 1, while Site 4 showed the lowest value. For both feature sets, the highest and lowest additive effects were found in Sites 3 and 1, respectively. As for GMV features, Site 3 was the site associated with highest. Among CT feature sets, Desikan-Killiany features showed higher than Destrieux ones. Consistently among GMV and CT features, the multiplicative coefficients were found to be maximum for Site 1, whereas additive coefficients were maximum for Site 3. Among all four feature sets, the highest sum of average additive and multiplicative coefficients across sites was found in Desikan-Killiany CT features. Within GMV and CT feature sets, differences were found between atlases and highlighted higher values for the CT features from the Desikan-Killiany atlas compared to CT features from the Destrieux one.

3.2 | ComBat Performance Test Removal of Sites Effects

3.2.1 | Feature Comparison Among Sites

After multiple comparison correction applied on KW analyses, we showed differences among sites in the median values of non-harmonized GMV and CT features, in both training ($p < 0.001$, surviving to Bonferroni's correction) and test ($p < 0.005$, surviving to Bonferroni's correction) sets. The post hoc pairwise comparisons showed that, for both training and test sets, the median GMV and CT values in Site3 were significantly lower than in the other sites. After ComBat harmonization, for both training and test sets, no significant differences in median feature values among sites were found for any of the GMV or CT feature sets.

3.2.2 | Visual Inspection of PCs

The scatterplot of the first two PCs related to GMV and CT features before and after ComBat were extracted. Scatterplots for GMV and CT training and test feature set is represented in Figures S2 and S3. The PCs are colored by site, enabling the visualization of their associations. Before ComBat harmonization, training set PCs were more clearly clustered by site compared to test set ones, probably due to the lower number of test set subjects from each site. Such clustering is not observable after ComBat harmonization.

3.2.3 | Site Classification Model

The accuracies of site classification models based on GMV and CT features before and after ComBat harmonization are reported in Table 3, for the training set, and Table 4, for the test set. Specifically, measures of BA, CA, specificity, sensitivity, and precision values reported in Tables 3 and 4 are calculated as the average value across folds. The BA and CA values were higher for the site classification models built on CT features compared to GMV features. Specifically, in the training set, the BA values of site classification based on Cobra and Neuromorphometrics GMV features were found to decrease of 65.17% and 80.25%, respectively, after ComBat harmonization, whereas classification models built on Destrieux and Desikan-Killiany CT features showed a reduction of BA values of 90.75% and 82.75%, respectively, after ComBat application. In the test set, ComBat application resulted in a BA reduction of 55.83% for Cobra GMV features, 26.25% for Neuromorphometrics GMV features, 48.33% for Destrieux CT features, and 47.8% for Desikan-Killiany CT features. In the training set, the models based on CT features were associated

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | Training set MCGP accuracies for GMV and CT features.

Not Harmonized	GMV Cobra	GMV Neuromorphometrics	CT Destrieux	CT Desikan-Killiany
BA	79.67%	90.58%	97.17%	95.25%
Sensitivity	79.55%	90.54%	97.32%	95.12%
Specificity	93.02%	96.82%	99.09%	98.33%
Precision	79.56%	90.55%	97.26%	95.23%
CA	CA1: 71.33%	CA1: 87%	CA1: 96%	CA1: 92.67%
	CA2: 80%	CA2: 87.67%	CA2: 100%	CA2: 98.33%
	CA3: 94%	CA3: 96%	CA3: 96%	CA3: 100%
	CA4: 73.33%	CA4: 91.67%	CA4: 96.67%	CA4: 90%
BA *p*-value permutation test	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Harmonized	**GMV Cobra**	**GMV Neuromorphometrics**	**CT Destrieux**	**CT Desikan-Killiany**
BA	14.50%	10.33%	6.42%	12.50%
Sensitivity	12.64%	9.15%	6.44%	11.59%
Specificity	71.67%	70.12%	68.67%	70.86%
Precision	14.42%	10.18%	6.51%	12.51%
CA	CA1: 2%	CA1: 0%	CA1: 8.33%	CA1: 23.33%
	CA2: 18.33%	CA2: 13.33%	CA2: 7.67%	CA2: 13%
	CA3: 6%	CA3: 8%	CA3: 8%	CA3: 2%
	CA4: 31.67%	CA4: 20%	CA4: 1.67%	CA4: 11.67%
BA *p*-value permutation	NS	NS	NS	NS
Accuracy Drop Before—After Harmonization	**GMV Cobra**	**GMV Neuromorphometrics**	**CT Destrieux**	**CT Desikan-Killiany**
Drop in BA	65.17%	80.25%	90.75%	82.75%
Drop in CA	CA1: 69.33%	CA1: 87%	CA1: 87.67%	CA1: 69.34%
	CA2: 61.67%	CA2: 74.34%	CA2: 92.33%	CA2: 85.33%
	CA3: 88%	CA3: 88%	CA3: 88%	CA3: 98%
	CA4: 41.66%	CA4: 71.67%	CA4: 95%	CA4: 78.33%

Abbreviations: BA, balanced accuracies; CA, class- (i.e., site-) and specific accuracies.

with the highest drop in BA values from before to after ComBat harmonization. Differently, in the test set, the highest BA drop was observed in the classification models built on Cobra GMV features, followed by the ones relying on CT features. By evaluating $BA_{train} - BA_{test}$, the lowest difference was found for Cobra GMV features (9.34%) followed by Desikan-Killiany CT (35.67%), Destrieux CT (42.42%), and GMV Neuromorphometrics (54%) ones. On average, such difference was greater for GMV features (31.67%) compared to CT ones (39.04%), indicating better ComBat performances on independent test sets for CT features than for GMV ones. In addition, Site 3 was associated with the highest drop of CA values, on average across the four feature sets, in both training (90%) and test (64%) sets.

Figures 2 and 3 show the histograms of distributions of average BA across folds and across permutations, before and after ComBat harmonization, as violin plots. Specifically, the left side of the violin plots represents the BA distribution of the model (colored in green in the "before harmonization" panels and dark pink in the "after harmonization" panels), while the right side corresponds to the distribution of average BA with permuted labels (colored in gray in the "before harmonization" panels, and purple in the "after harmonization" panels). Furthermore, as a summary measure across CV folds, the multiclass confusion matrices for the training and the test set are reported in Figures S4 and S5, representing the average confusion matrix across CV folds matrices.

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Test set MCGP accuracies for GMV and CT features.

Not Harmonized	GMV Cobra	GMV Neuromorphometrics	CT Destrieux	CT Desikan-Killiany
BA	73.75%	75%	88.75%	86.25
Sensitivity	81.54%	78.54%	92.01%	95.12%
Specificity	93.95%	92.60%	97.30%	98.33%
Precision	81.35%	77.87%	92.16%	95.23%
CA	CA1: 72.22%	CA1: 77.78%	CA1: 94.44%	CA1: 77.78%
	CA2: 63.16%	CA2: 73.68%	CA2: 84.21%	CA2: 94.74%
	CA3: 100%	**CA3: 100%**	**CA3: 100%**	**CA3: 100%**
	CA4: 90%	CA4: 60%	CA4: 90%	CA4: 85%
BA *p*-value permutation	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.05$
Harmonized	**GMV Cobra**	**GMV Neuromorphometrics**	**CT Destrieux**	**CT Desikan-Killiany**
BA	17.92%	48.75%	40.42%	39.17%
Sensitivity	12.64%	50.44%	42.25%	41.54%
Specificity	71.67%	83.27%	80.82%	80.34%
Precision	14.42%	50.08%	41.92%	40.97%
Site's specific accuracy	CA1: 22.22%	CA1: 55.56%	CA1: 33.33%	CA1: 44.44%
	CA2: 21.05%	CA2: 36.84%	CA2: 63.16%	CA2: 47.37%
	CA3: 0%	CA3: 52.94%	CA3: 41.18%	CA3: 47.06%
	CA4: 30%	CA4: 55%	CA4: 30%	CA4: 25%
BA *p*-value permutation	NS	$p < 0.05$	$p < 0.05$	$p < 0.05$
Accuracy Drop Before–After Harmonization	**GMV Cobra**	**GMV Neuromorphometrics**	**CT Destrieux**	**CT Desikan-Killiany**
Drop in BA	55.83%	26.25%	48.33%	47.08%
Drop in CA	CA1: 50%	CA1: 22.22%	CA1: 61.11%	CA1: 33.34%
	CA2:42.11%	CA2:36.84%	CA2: 21.05%	CA2: 47.37%
	CA3: 100%	CA3: 47.06%	CA3: 58.82%	CA3: 52.94%
	CA4:60%	CA4:5%	CA4:60%	CA4: 60%

Abbreviations: BA, balanced accuracies; CA, class- (i.e., site-) and specific accuracies.

In the training set, before ComBat harmonization, significant *p*-values from the permutation tests were found for the BA values for all feature sets, indicating the rejection of the null hypothesis of independence between samples (i.e., GMV and CT features) and classes (i.e., sites). After harmonization, the same BA metrics were associated with nonsignificant *p*-values from the permutation tests.

In the test set, before ComBat harmonization, the permutation tests resulted in significant *p*-values for BA values for all feature sets. Differently, after harmonization, only GMV Cobra features resulted in nonsignificant BA differences between permuted and true labels. Differently, GMV Neuromorphometrics, Destrieux and Desikan-Killiany CT feature sets were associated with significant BA differences.

3.3 | Preservation of Biological Variability

3.3.1 | Linear Regression Analysis

The linear regression statistics relative to the associations between biological factors (age and sex) and median feature values are reported in Table 5, for the training set, and Table 6, for the test set. Overall, ComBat harmonization strengthened the association of age and sex with the median GMV and CT features in

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the training set, whereas this association was reinforced only for GMV features in the test set. In the training set, for all feature sets, the R^2 coefficient from the linear regression model increased after harmonization, indicating a post-harmonization increase in the percentage of variation in the response variable (median feature value) explained by age and sex. On the other hand, in the test set, the R^2 coefficient increased from before to after harmonization only for GMV features, whereas a slight decrease was observed for the CT features.

Regarding age, in the training set, for both GMV and CT features, a significant association between age and median feature values was maintained from before to after ComBat harmonization; moreover, the absolute value of the T-statistics relative to the age regressor increased after harmonization. In the test set, the significant associations between age and non-harmonized median feature values, which was observed for the GMV Neuromorphometrics and CT Destrieux feature sets, were maintained after harmonization. In the same line, nonsignificant associations observed for the GMV Cobra and CT Desikan-Killiany feature sets before harmonization remained nonsignificant after ComBat harmonization. Regarding the sex factor, in the training set, the magnitude of its association with median GMV feature values (i.e., the T-statistics) increased after harmonization, maintaining a significant *p*-value. Differently, median CT features showed a nonsignificant association with sex both before and after harmonization. As in the training set, in the test set the strength of the association between sex and median GMV features from both Cobra and Neuromorphometrics increased from before to after harmonization, maintaining a significant *p*-value. The median CT features were not significantly associated with sex both before and after ComBat harmonization. Regarding the linear regression analyses assessing the associations between individual GMV and CT features and biological factors, we found that, overall, in both training and test sets, after ComBat harmonization, an increased number of ROIs were significantly associated with age and sex for most of the feature sets, except for CT Desikan-Killiany atlas with sex in the training set and GMV Neuromorphometrics atlas with age in the test set.

The findings on the individual features are summarized in terms of numerosity of significant associations before and after ComBat harmonization. In the training set, before harmonization, we found 30 GMV Cobra, 115 GMV Neuromorphometrics, 121 CT Destrieux, and 59 Desikan-Killiany features significantly associated with age, and 39 GMV Cobra, 118 GMV Neuromorphometrics, 32 Destrieux, and 16 Desikan-Killiany features significantly associated with sex. After harmonization, the number of features significantly associated with sex increased for all feature sets except for CT Destrieux one, whereas the number of features significantly associated with age did not increases after harmonization. Specifically, 27 GMV Cobra, 112 GMV Neuromorphometrics, 120 CT Destrieux, and 58 Desikan-Killiany features were significantly associated with age, and 40 GMV Cobra, 120 GMV Neuromorphometrics, 31 Destrieux, and 16 Desikan-Killiany features were significantly associated with sex. In the test set, before harmonization, we assessed that none of GMV Cobra, 62 GMV Neuromorphometrics, 69 CT Destrieux and 35 Desikan-Killiany features were significantly associated with age. Further, 33 GMV Cobra, 103 GMV Neuromorphometrics, 4 CT Destrieux and 1 Desikan-Killiany regions had a significant association with sex. Of note, differently from the training set, the number of regions significantly associated with age after the test set harmonization increased for all sets except for Neuromorphometrics atlas, whereas the regions associated with sex after test set harmonization increased for all the GMV and CT atlases. Specifically, we found 2 GMV Cobra, 59 GMV Neuromorphometrics, 74 CT Destrieux, 36 CT Desikan-Killiany features significantly associated with age after harmonization, while 37 GMV Cobra, 109 GMV Neuromorphometrics, 12 CT Destrieux, and 6 CT Desikan-Killiany were significantly associated with sex.

3.3.2 | KRR for Age

The results of KRR application for prediction of age based on GMV and CT features from the training and test sets are reported in Tables S5 and S6, respectively. In the training set, for all feature sets, the KRR performed better after ComBat harmonization compared to before harmonization, as indicated by a decrease of MSE from before to after ComBat harmonization for all models ($\Delta_{MSE\ pre-post\ harm}$: GMV Cobra, 2.97; GMV Neuromorphometrics, 7.76; CT Destrieux, 8.51; CT Desikan-Killiany, 6.91). On the other hand, for the test set, ComBat harmonization improved the KRR prediction accuracy (i.e., decreasing MSE) of age in the models built on CT features, whereas the age prediction performances remained unchanged in the models built on GMV features from before to after Combat harmonization ($\Delta_{MSE\ pre-post\ harm}$: GMV Cobra, 0; GMV Neuromorphometrics, 5.05; CT Destrieux, 0.44; CT Desikan-Killiany, 7.29).

3.3.3 | Binary SVM for Sex Classification

In the training set, ComBat harmonization improved the sex classification accuracy only when GMV features were used, whereas for the models relying on CT features the performance decreased after ComBat harmonization. This result has confirmed that the biological variability associated with sex was preserved, and even strengthened, in the harmonized GMV features ($\Delta_{Accuracy\ drops\ post-pre\ harm}$: GMV Cobra, 4.82; GMV Neuromorphometrics, 2.68; CT Destrieux, −3.32; CT Desikan-Killiany, −8.5). In the test set, the sex classification performances increased when harmonized GMV features compared to non-harmonized GMV features were used. Differently from the training set, ComBat improved the sex classification accuracy in the models relying on CT Destrieux features. As in the training set, a drop in sex classification accuracy from before to after ComBat harmonization was observed for CT Desikan-Killiany features ($\Delta_{Accuracy\ drops\ post-pre\ harm}$: GMV Cobra, 1.67; GMV Neuromorphometrics, 10.41; CT Destrieux, 12.92; CT Desikan-Killiany, −2.92).

4 | Discussion

Neuroscientific advances have recently brought to the generation of large amounts of neuroimaging data that are increasingly

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

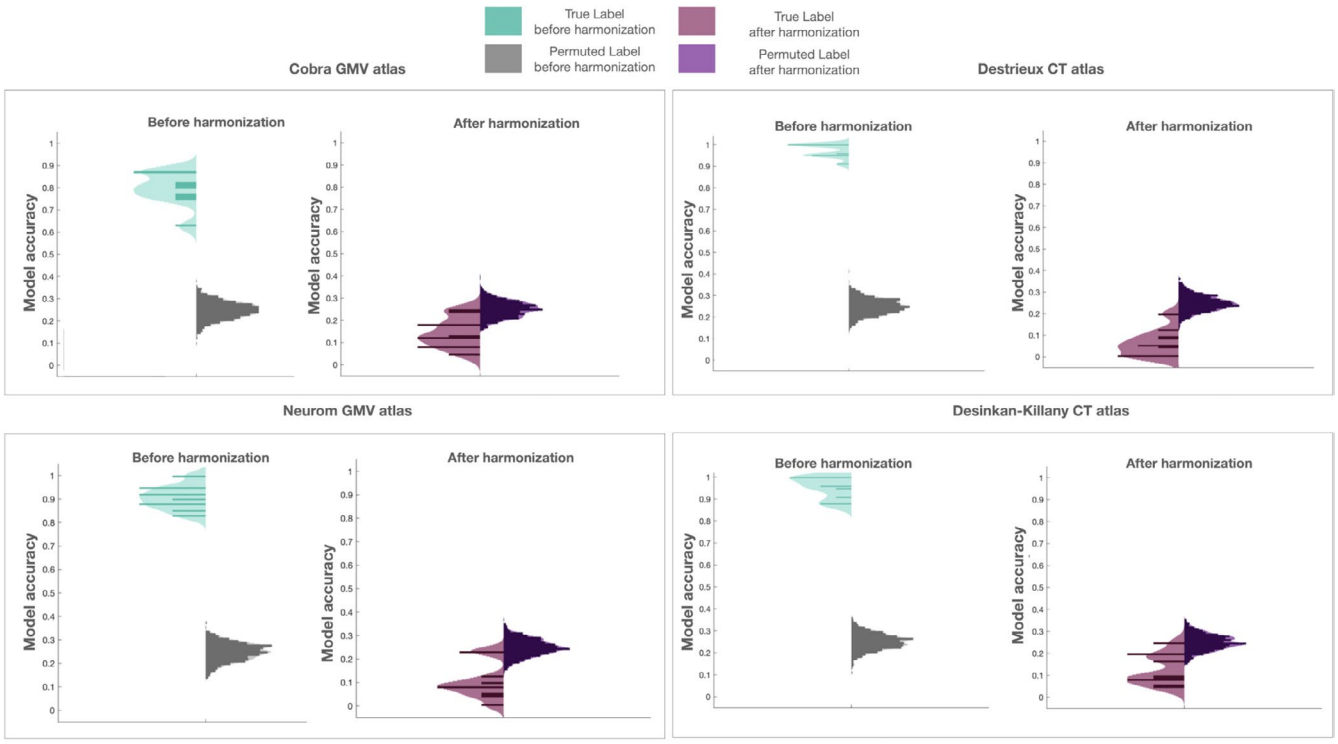

FIGURE 2 | Histogram of distribution of average balanced accuracy across folds and across permutations as a violin plot in training set, reported before and after ComBat harmonization. The left side of each violin plots (colored in green before harmonization and dark pink after harmonization) represents the balanced accuracy distribution of the model, while the right side of each plots (colored in gray before harmonization and purple after harmonization) corresponds to the distribution of average balanced accuracy with permuted labels.

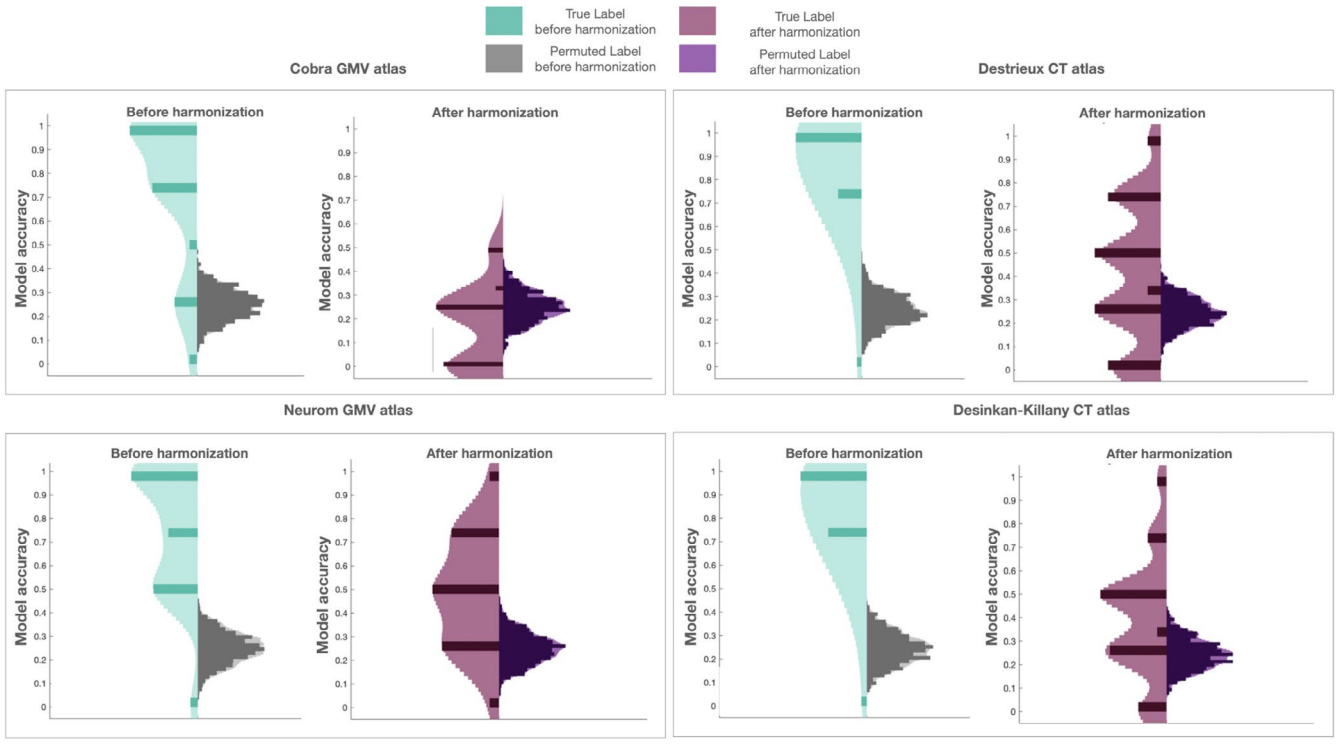

FIGURE 3 | Histogram of distribution of average balanced accuracy across folds and across permutations as a violin plot in test set, reported before and after ComBat harmonization. The left side of each violin plots (colored in green before harmonization and dark pink after harmonization) represents the balanced accuracy distribution of the model, while the right side of each plots (colored in gray before harmonization and purple after harmonization) corresponds to the distribution of average balanced accuracy with permuted labels.

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 | Statistical results from linear regression model applied on training set with dependent variables of age and sex and median GMV and CT measures as outcome.

	GMV Cobra		GMV Neurom		CT Destrieux		CT Desikan-Killiany	
Age	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**
tStats	−4.87	−5.03	−6.59	−6.66	−7.46	−8.09	−7.77	−8.25
p value	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$
	GMV Cobra		**GMV Neurom**		**CT Destrieux**		**CT Desikan-Killiany**	
Sex	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**
tStats	6.71	8.36	8.21	10.69	−1.19	−0.78	−1.26	−0.98
p value	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$	NS	NS	NS	NS
R^2	23.8%	30.3%	33.6%	42%	20.9%	23.4%	22.3%	24.2%

TABLE 6 | statistical results from linear regression model applied on test set with dependent variables of age and sex and median GMV and CT measures as outcome.

	GMV Cobra		GMV Neurom		CT Destrieux		CT Desikan-Killiany	
Age	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**
tStats	−1.08	−1.28	−3.05	−3.06	−3.07	−3.05	−2.88	−2.83
p value	NS	NS	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$	NS	NS
	GMV Cobra		**GMV Neurom**		**CT Destrieux**		**CT Desikan-Killiany**	
Sex	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**	**PreHarm**	**PostHarm**
tStats	4.89	5.61	3.55	4.08	−0.78	−0.55	−0.70	−0.67
p value	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$	NS	NS	NS	NS
R^2	29.30%	35.41%	29%	32.50%	11.70%	11.60%	10.51%	10.10%

shared among centers to respond to a range of research questions. Within this context, the ComBat tool is gaining importance as harmonization technique for removing non-biological variability in neuroimaging data, such as the one induced by site-specific acquisition parameters.

In this study, we applied ComBat to multisite sMRI data including diverse brain regional types of features and diverse types of atlases per feature via a CV framework, and quantitatively tested its performance by comparing harmonized and non-harmonized features. For the first time, surface-based and volume-based measures of regional brain morphology from multiple atlases were considered, and multiple performance metrics relative to both non-biological (site-related) and biological variability in the data were extracted, enabling an extensive characterization of the ComBat harmonization pipeline. Of note, in our study, ComBat harmonization effectiveness was tested in a sample that is representative of most multisite samples investigated in literature.

Specifically, besides the use of standard statistical tests to evaluate Combat ability to remove site-related differences in features, ML classification based on MCGP was applied. By employing a classification model to estimate site effect removal, we ensured a quantitative assessment of non-biological site-related variability based on various performance metrics, including accuracy and sensitivity, further assessing the robustness of these metrics through permutation tests. Additionally, using ML classification models, the robustness of harmonization on independent sets was derived, as well as the best predictive features for the site's classification.

Our results support the flexibility and robustness of ComBat across multiple surface- and volume-based measures of brain morphology, valuing its usefulness in removing site effect also on independent data from the same sites and in preserving biological variability across participants in the data. Furthermore, differences in terms of ComBat performances were found to varied across measures of regional brain morphology, and thus demonstrating tissue specific site effects' modeling. Moreover, ComBat adjustment of site effects also varied across the regional level of each specific volume-based and surface-based measures. Overall, so far, few studies quantitatively evaluated ComBat efficacy on MRI data, and among them, even fewer provided statistical metrics of harmonization performances. Of note, only one study focused on sMRI features (Yu et al. 2018; Fortin et al. 2018; Yamashita et al. 2019; Richter et al. 2022; Maikusa et al. 2021). Specifically, they employed statistical analysis and unsupervised dimensionality reduction to relatively test for sites effects and to visually explore if data's variation remained associated

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with sites after harmonization. Similarly, recent fMRI studies applied statistical analysis (Yu et al. 2018), and unsupervised clustering methods (Yamashita et al. 2019) as well as a multi-scale assessment of harmonization efficacy to evaluate a specific quality control metric (Tassi et al. 2023). Nevertheless, the application of ComBat to the entire dataset has impeded the external validation of site classification accuracy and the assessment of result reproducibility, specifically for cross-validated ComBat application. To our knowledge, our study is the first to assess the flexibility and robustness of ComBat in harmonizing independent test sets, providing supportive evidence that is relevant for many ML applications.

4.1 | Multiplicative and Additive Site Effects Analysis

In the panorama of ComBat applications (Pomponio et al. 2020; Fortin et al. 2017; Tassi et al. 2023; Fortin et al. 2018; Beer et al. 2020; Radua et al. 2020; Yamashita et al. 2019; Chen 2022), our study is innovative for focusing on the separate analysis of additive and multiplicative coefficient differences across sites. Within our application, we found an overall homogeneity of these effects across surface- and volume-based measures, but with heterogeneous distributions of multiplicative coefficients with respect to additive ones. Specifically, Sites 3 and 1 were associated with the highest multiplicative and additive coefficients, respectively, consistently across GMV and CT atlases. Thus, our evidence suggests that site-related variability in sMRI data are reflected evenly in both volume-based and surface-based feature sets.

4.2 | Site Effects Removal

For the first time, we employed the MCGP for site classification based on both non-harmonized and harmonized features for quantifying ComBat performances. The comparison of site classification accuracy metrics between training and test sets and across feature sets provided novel information on ComBat performances, especially in the context of sMRI studies.

In terms of BA, our results showed a drop from before to after ComBat application for both GMV and CT features, with the higher drop observed on Destrieux CT features for the training set, on Cobra GMV features for test set. The permutation test results showed that ComBat harmonization on training feature sets removed the non-biological site-related variability, and hence the association between labels (i.e., imaging sites) and features. Otherwise, ComBat was found to be more effective in removing site-to-site variation from independent (test) samples for volume-based. Accordingly, the performance differences from training to test set were on average bigger for GMV features compared to CT ones. Furthermore, the comparison among sites in terms of drop in CA after harmonization led to results that are consistent with the inter-site differences in multiplicative and additive effects. Specifically, in both training and test sets, the site associated with highest multiplicative coefficients was also characterized by the highest CA drop on average across GMV and CT features. These results also suggest that scale differences across sites, modeled through multiplicative coefficients, were determinant for site classification. Overall, our findings suggest that the inclusion of ComBat in a CV framework might leave some of the undesired site-related variability in brain morphological features, especially those extracted using surface-based approaches. These results have highlighted the importance of analyzing ComBat harmonization performances on diverse types of measurements, including both volume-based and surface-based features, as well as diverse atlases for each of them. Therefore, we introduced the need to carefully consider the dataset properties before applying the method since not all feature types might be suitable for ComBat application in CV frameworks. The different performances observed in the test sets are probably due to the inclusion of ComBat in the CV framework. Since ComBat takes into account the effects of the biological covariates estimated on the training set, its application to the test set might have implied a less accurate consideration of biological variability, affecting in turn the magnitude of association between covariates and harmonized features. The observed discrepancies between GMV and CT features might also result from differences in their association with age. ComBat performances might thus vary across feature sets due to differences in the relative relationships with biological features, remarking the need to develop customized pipelines. Interestingly, previous studies have reported dynamic changes in CT, volume and surface area throughout the adult life span, underlying vulnerability of specific regions to age-related modifications (Storsve 2014). However, the nonlinearity of age trends in a wide number of brain features seems to be reduced in adulthood compared to adolescence (Pomponio et al. 2020). Thus, in our adult sample, we could assume discardable nonlinear age-related modifications in GMV and CT features. Differently, the results of the linear regression models for the variable of sex showed that, in the training set, ComBat strengthened the relationship between GMV features and sex, while the strength of the association of sex with CT measures decreased after the harmonization. Nevertheless, in the independent test set, the association strength increased for all the feature sets except for CT Desikan-Killiany ones. Likely, previous findings on CT data have shown no effects of ComBat harmonization on the association between features and sex (Fortin et al. 2018). Furthermore, results related to ML classification of sex in the training set were consistent with linear regression ones, showing that harmonization improved the classification of sex from GMV features. Given the paucity of studies addressing this issue, our findings need replication on independent sets before drawing firm conclusions on the impact of ComBat on the relationships between brain morphological features and sex.

Therefore, our quantification of site effect removal on unseen independent datasets from the same site confirms ComBat flexibility and robustness on independent gray matter data, highlighting its potential for broader applications. Thus, it would facilitate screening deviations effects in independent datasets, without the need of re-deriving specific site effects' coefficients.

5 | Limitations

Potential limitations related to the application of ComBat may arise and need to be discussed. As first ComBat technique relies

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

on strong assumptions for the parametric prior distributions of site effects, assuming that multiplicative and additive effects follow normal and inverse gamma distributions, respectively. Such parametric prior distributions might not generalize to all possible scenarios for input features. Moreover, this hypothesis poses three main implications. As first implication, ComBat does not account for heteroscedastic distributions of target features. Thus, it considers the features drawn from the same distribution with single mean and variance for each site and furthermore forces all sites to align with a single homoscedastic noise term distribution. As second implication, the standard deviation of target features is considered as constant, loosing important biological meaning in the data that may exist by considering heteroscedastic features' distributions. Finally, the adjustment of site effects might be unbalanced in the case of large differences in the sample size across sites, leading a heavily adjustment for the sites with less data with respect to sites with higher sample size. Nevertheless, ComBat is funded on an empirical Bayes step, which ensures robustness toward outliers and solid performances in small-to-large samples (Johnson, Li, and Rabinovic 2007; Fortin et al. 2018; Radua et al. 2020; Bayer 2022; Chen et al. 2011). Although the performance of ComBat has been demonstrated to be stable across different samples (with $N > 25$) (Fortin et al. 2017; Bayer 2022; Chen et al. 2011), future research could extend our sample in terms of numerosity and age range covered and employ subsampling strategies to verify the effect created by the training sample size on the site effect estimation, thereby clarifying the extent to which the sample size could influence the harmonization within an independent test set.

However, our training and test sets were balanced among sites, guaranteeing a robust batch effect adjustment in this context. Nonetheless, besides implications ComBat have demonstrated robustness and widely reliable results in harmonizing ROI-based neuroimaging features, as well as voxel-based (Pomponio et al. 2020; Fortin et al. 2018; Beer et al. 2020; Chen 2022). Within this context, alternative prior distributions' set-up, as the use of non-parametric hierarchical Bayesian priors in the batch effect parameters should be applied to evaluate the differences in ComBat harmonization performances based on different priors' distribution. However, fully non-parametric Bayesian approach may require approximation of too complex and high-dimensional functions. Nevertheless, each Bayesian design suffers from specific issues, thus comparative analysis on multiplicative and additive site effects based on parametric and nonparametric prior are needed.

Furthermore, in this study we compared across features and sites the total average magnitude of site's estimators (i.e., $\delta_{mean} + \gamma_{mean}$), identified based on a composite index indicative of the sum of the magnitudes related to multiplicative and additive site effects. Although the composite index of the overall site effect represents a single metric that helped us to localize, evaluate and compare the summary of multiplicative and additive sources among feature sets and sites, we considered it a "suboptimal" indicator of the overall site effects estimated by ComBat, thus considering that the magnitudes of both site effects sources are not easily integrable. Indeed, the correct removal of the overall inter-site variability relies on the ComBat equations, taking into account diverse factors including the multiplicative and additive effects, as well as covariates and error terms.

Moreover, our sample of young adults limited the possibility of further age sensitivity analyses, such as the assessment of biological factors' preservation over development. Even if our dataset covered a relatively wide age range, subjects covering a life span, from childhood to elderhood, were absent. Future analyses should include the investigation of ComBat capability to preserve age effects on lifespan samples and subsamples belonging to different age sections.

In addition, our analysis may be limited by the specific division of the data in training and test set. Alternative analyses based on multiple iterations in the splitting process might be applied to enhance reliability of ComBat performances on training and test sets. Further, in this study, we ensured sample sizes and biological covariates balanced across sites to perform as best the aim of quantitative evaluation of the ComBat harmonization effectiveness. However, our approach might not be representative of practical applications characterized by imbalance among sites in sample sizes and covariates. Although the robustness of ComBat demonstrated toward imbalanced conditions in literature (Johnson, Li, and Rabinovic 2007; Fortin et al. 2018), within an imbalanced dataset the correct harmonization and maintenance of biological factors could not be guaranteed. Mitigation strategies could be applied, including the estimation of site effects parameters within balanced subsamples from the different sites (Gupta et al. 2022).

Another limitation of ComBat is the assumption of independency of the biological effects with respect to the site effects. Although ComBat ensures to preserve biological effects accounted for in the algorithm, any kind of multicollinearity between sites and covariates is not modeled thus leading ComBat to over- or under-correct the feature values of the site with this collinearity by forcing them to align with a linear covariate trajectory. Nevertheless, by considering a wide young adults' age range we might exclude heavy over and under adjustments due to large nonlinear covariate effects. Making a specific focus on the risk for over-correction, we considered limited the risk of removing effects of interest, such as the ones covarying with site, since the removal of site-related heterogeneity was in most cases generalizable to the multisite independent test set, especially for GMV features, showing nonsignificant BA differences between permuted and true labels. A further issue to be considered is related to the repetition of site labels entered in the permutation framework. However, over-correction's risk within ComBat framework needs to be validated on independent samples with heterogeneous characteristics across sites, as well as how much this over adjustment could interfere with the effects of interest estimation.

The nonsignificant differences age ranges across sites in our study should have guaranteed a more stable batch effect correction with respect to age-disjoint studies (Pomponio et al. 2020; Bayer 2022). Thus, these results underlined the importance of considering possible unmodeled site-by-covariate interactions acting on feature sets in the context of ComBat application. The main limitation of our study regards the non-consideration of harmonization methods different from

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ComBat one. Alternative methods, including Combat variants, should be applied in the future. Among them, ComBat-GAM has been introduced for modeling nonlinear covariate effects on input features (Pomponio et al. 2020) ComBat provides harmonization of site-specific covariance patterns across features (Chen 2022) and ComBat longitudinal approach can be applied on repeated measurements of the same subjects (Beer et al. 2020).

Future studies would benefit from replicate these findings on larger sample size, wider age range and more sites involved. Thus, replication of our findings on larger datasets would provide a quantitative advantage on ComBat performance's estimation leading to a more solid and robust estimate of tissue specific site effects, as well as an opportunity to verify flexibility, robustness, and generalizability on unseen dataset.

Further, next critical steps should provide more detailed information on ComBat performances in terms of maintenance of biological variability within single GMV and CT features. This would ensure the assessment not only across multiple types of features, but also at finer spatial scales, providing further information at the ROI and sub-ROI levels. However, looking at which brain regions are more affected by ComBat could be misleading, as these brain regions can vary greatly depending on the dataset used.

As last, the employment of traveling subjects' design may represent a valuable future validation of ComBat performances, taking into account also the site-specific samples' effect.

6 | Conclusion

Our study has quantitatively assessed ComBat multisite harmonization performance on different types of regional brain morphology metrics via a novel CV approach. Using a robust ML and statistical tools, we found that ComBat is effective at removing nuisance variability in the data associated with site, while generally preserving or even strengthening data association with biological factors. In our application, especially for GMV features, ComBat has shown flexibility and robustness of application on unseen independent GMV data from the same sites. If reproduced on larger independent samples, this evidence might provide a conclusive insight into ComBat effectiveness in the context of ML studies on regional brain morphological features. Thus, especially the results of ComBat CV strategy may have considerable implications in the field of developing classification tools for psychopathological disorders based on large multisite clinical datasets, remarking the importance of transfer the site-specific coefficients on unseen data from the same sites.

Author Contributions

E.T.: data curation; methodology, software, writing – original draft, writing – review and editing. A.M.B.: methodology review, writing – review and editing. F.C.: methodology review, writing – review and editing. B.V.: methodology review, writing – review and editing. M.B.: data curation, writing – review and editing. I.N.: data curation, writing – review and editing. F.P.: data curation, writing – review and editing. F.B.: data curation, writing – review and editing. P.B.: data curation, conceptualization, project administration, supervision, writing – review and editing. E.M.: data curation; methodology review; supervision; writing – original draft.

Acknowledgments

Open access funding provided by BIBLIOSAN.

Ethics Statement

This study protocol received the approval of Ethics Committee of the four sites of the StratiBip network, which stemmed from the ENPACT network (Delvecchio et al. 2021, Maggioni et al. 2017).

Consent

All participants provided a written consent to the study protocol.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Demographic and MRI data underlying the findings presented in this manuscript are protected by privacy and will be available from the corresponding author upon reasonable request, after signature of a formal data sharing agreement. MATLAB codes are available at Opens Science Framework web application link: https://osf.io/ksmja/?view_only=bddd0aa06d26487fac7551f6e3109a1f.

References

Bayer, J. M. M. 2022. "Site Effects How-To and When: An Overview of Retrospective Techniques to Accommodate Site Effects in Multi-Site Neuroimaging Analyses." *Frontiers in Neurology* 13: 923988. https://doi.org/10.3389/fneur.2022.923988.

Beer, J. C., N. J. Tustison, P. A. Cook, et al. 2020. "Longitudinal ComBat: A Method for Harmonizing Longitu- Dinal Multi-Scanner Imaging Data." *NeuroImage* 220: 117129. https://doi.org/10.1016/j.neuroimage.

Chen, A. A. 2022. "Mitigating Site Effects in Covariance for Machine Learning in Neuroimaging Data." *Human Brain Mapping* 43: 1179–1195.

Chen, C., K. Grennan, J. Badner, et al. 2011. "Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods." *PLoS One* 6, no. 2: e17238. https://doi.org/10.1371/journal.pone.0017238.

Delvecchio, G., E. Maggioni, A. Pigoni, et al. 2021. "Sexual Regional Dimorphism of Post-Adolescent and Middle Age Brain Maturation. A Multi-Center 3T MRI Study." *Frontiers in Aging Neuroscience* 13: 622054. https://doi.org/10.3389/fnagi.2021.622054.

Desikan, R. S. 2006. "An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans Into Gyral Based Regions of Interest." *NeuroImage* 31: 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021.

Destrieux, C., B. Fischl, A. Dale, and E. Halgren. 2010. "Automatic Parcellation of Human Cortical Gyri and Sulci Using Standard Anatomical Nomenclature." *NeuroImage* 53: 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010.

Fortin, J. P., N. Cullen, Y. I. Sheline, et al. 2018. "Harmonization of Cortical Thickness Measurements Across Scanners and Sites." *NeuroImage* 167: 104–120. https://doi.org/10.1016/j.neuroimage.

Fortin, J.-P., D. Parker, B. Tunç, et al. 2017. "Harmonization of Multi-Site Diffusion Tensor Imaging Data." *NeuroImage* 161: 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047.

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Gupta, R., S. Abdalla, V. Meausoone, et al. 2022. "Effect of Imbalanced Sampling and Missing Data on Associations Between Gender Norms and Risk of Adolescent HIV." *eClinicalMedicine* 50: 101513. https://doi.org/10.1016/j.eclinm.2022.101513.

Johnson, W. E., C. Li, and A. Rabinovic. 2007. "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods." *Biostatistics* 8, no. 1: 118–127. https://doi.org/10.1093/biostatistics/kxj037.

Maggioni, E., B. Crespo-Facorro, I. Nenadic, et al. 2017. "Common and Distinct Structural Features of Schizophrenia and Bipolar Disorder: The European Network on Psychosis. Affective Disorders and Cognitive Trajectory (ENPACT) Study." *PLoS One* 12: e0188000.

Maikusa, N., Y. Zhu, A. Uematsu, et al. 2021. "Comparison of Traveling-Subject and ComBat Harmonization Methods for Assessing Structural Brain Characteristics." *Human Brain Mapping* 42, no. 16: 5278–5287. https://doi.org/10.1002/hbm.25615.

Park, M. T. M. 2014. "Derivation of High-Resolution MRI Atlases of the Human Cerebellum at 3T and Segmentation Using Multiple Automatically Generated Templates." *NeuroImage* 95: 217–231. https://doi.org/10.1016/j.neuroimage.2014.03.037.

Penny, W. D., K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. 2011. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Academic Press.

Pomponio, R., G. Erus, M. Habes, et al. 2020. "Harmonization of Large MRI Datasets for the Analysis of Brain Imaging Patterns Throughout the Lifespan." *NeuroImage* 208: 116450. https://doi.org/10.1016/j.neuroimage.2019.116450.

Radua, J., E. Vieta, R. Shinohara, et al. 2020. "Increased Power by Harmonizing Structural MRI Site Differences With the ComBat Batch Adjustment Method in ENIGMA." *NeuroImage* 218: 116956. https://doi.org/10.1016/j.neuroimage.2020.116956.

Rasmussen, C., and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Richter, S., S. Winzeck, M. M. Correia, et al. 2022. "Validation of Cross-Sectional and Longitudinal ComBat Harmonization Methods for Magnetic Resonance Imaging Data on a Travelling Subject Cohort." *NeuroImage: Reports* 2, no. 4: 100136. https://doi.org/10.1016/j.ynirp.2022.100136.

Schrouff, J. 2013. "Pattern Recognition for Neuroimaging Toolbox." *Neuroinformatics* 11: 319–337. https://doi.org/10.1007/s12021-013-9178-1.

Storsve, A. B. 2014. "Differential Longitudinal Changes in Cortical Thickness, Surface Area and Volume Across the Adult Life Span: Regions of Accelerating and Decelerating Change." *Journal of Neuroscience* 34: 8488–8498. https://doi.org/10.1523/JNEUROSCI.0391-14.2014.

Takao, H., N. Hayashi, and K. Ohtomo. 2011. "Effect of Scanner in Longitudinal Studies of Brain Volume Changes." *Journal of Magnetic Resonance Imaging* 34, no. 2: 438–444. https://doi.org/10.1002/jmri.22636.

Tassi, E., F. Goffi, M. G. Rossetti, et al. 2023. "Multi-Scale Assessment of Harmonization Efficacy on Resting-State Functional Connectivity." *MEDICON'23 & CMBEBIH'23 Proceeding* 93: 301–308. https://doi.org/10.1007/978-3-031-49062-0_33.

Wegrzyn, M., M. Riehle, K. Labudda, et al. 2015. "Investigating the Brain Basis of Facial Expression Perception Using Multi-Voxel Pattern Analysis." *Cortex* 69: 131–140. https://doi.org/10.1016/j.cortex.2015.05.003.

Yamashita, A., N. Yahata, T. Itahashi, et al. 2019. "Harmonization of Resting-State Functional MRI Data Across Multiple Imaging Sites via the Separation of Site Differences Into Sampling Bias and Measurement Bias." *PLoS Biology* 17, no. 4: e3000042. https://doi.org/10.1371/journal.pbio.3000042.

Yu, M., K. A. Linn, P. A. Cook, et al. 2018. "Statistical Harmonization Corrects Site Effects in Functional Connectivity Measurements From Multi-Site fMRI Data." *Human Brain Mapping* 39, no. 11: 4213–4227. https://doi.org/10.1002/hbm.24241.

Zavaliangos-Petropulu, A. 2019. "Diffusion MRI Indices and Their Relation to Cognitive Impairment in Brain Aging: The Updated Multi-Protocol Approach in ADNI3." *Frontiers in Neuroinformatics* 13: 2. https://doi.org/10.3389/fninf.2019.00002.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

10970193, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.70085 by IRCCS Ospedale San Raffaele, Wiley Online Library on [28/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License