Received: 1 August 2021 | Revised: 10 February 2022 | Accepted: 7 March 2022

DOI: 10.1002/hbm.25838

RESEARCH ARTICLE

WILEY

Parameters from site classification to harmonize MRI clinical studies: Application to a multi-site Parkinson's disease dataset

Gemma C. Monte-Rubio[1,2] | Barbara Segura[1,2,3,4] | Antonio P. Strafella[5,6,7] | Thilo van Eimeren[8,9] | Naroa Ibarretxe-Bilbao[10] | Maria Diez-Cirarda[7] | Carsten Eggers[11,12,13] | Olaia Lucas-Jiménez[10] | Natalia Ojeda[10] | Javier Peña[10] | Marina C. Ruppert[11,12] | Roser Sala-Llonch[1,3,14,15] | Hendrik Theis[9] | Carme Uribe[1,2,3,7] | Carme Junque[1,2,3,4]

[1]Institute of Neurosciences, University of Barcelona, Barcelona, Catalonia, Spain

[2]Medical Psychology Unit, Department of Medicine, University of Barcelona, Barcelona, Catalonia, Spain

[3]Institute of Biomedical Research August Pi i Sunyer (IDIBAPS), Barcelona, Catalonia, Spain

[4]Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED: CB06/05/0018-ISCIII) Barcelona, Barcelona, Catalonia, Spain

[5]Edmond J. Safra Parkinson Disease Program & Morton and Gloria Shulman Movement Disorder Unit, Neurology Division, University Health Network, University of Toronto, Toronto, Ontario, Canada

[6]Krembil Brain Institute, University Health Network, University of Toronto, Toronto, Ontario, Canada

[7]Brain Health Imaging Centre, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health University of Toronto, Toronto, Ontario, Canada

[8]Department of Nuclear Medicine, University of Cologne, Cologne, Germany

[9]Department of Neurology, University of Cologne, Cologne, Germany

[10]Department of Psychology, Faculty of Health Sciences, University of Deusto, Bilbao, Spain

[11]Department of Neurology, University Hospital Marburg, Marburg, Germany

[12]Center for Mind, Brain and Behavior – CMBB, Universities Marburg and Gießen, Marburg and Gießen, Germany

Abstract

Multi-site MRI datasets are crucial for big data research. However, neuroimaging studies must face the batch effect. Here, we propose an approach that uses the predictive probabilities provided by Gaussian processes (GPs) to harmonize clinical-based studies. A multi-site dataset of 216 Parkinson's disease (PD) patients and 87 healthy subjects (HS) was used. We performed a site GP classification using MRI data. The outcomes estimated from this classification, redefined like Weighted HARMonization PArameters (WHARMPA), were used as regressors in two different clinical studies: A PD versus HS machine learning classification using GP, and a VBM comparison (FWE-$p < .05$, $k = 100$). Same studies were also conducted using conventional Boolean site covariates, and without information about site belonging. The results from site GP classification provided high scores, balanced accuracy (BAC) was 98.39% for grey matter images. PD versus HS classification performed better when the WHARMPA were used to harmonize (BAC = 78.60%; AUC = 0.90) than when using the Boolean site information (BAC = 56.31%; AUC = 0.71) and without it (BAC = 57.22%; AUC = 0.73). The VBM analysis harmonized using WHARMPA provided larger and more statistically robust clusters in regions previously reported in PD than when the Boolean site covariates or no corrections were added to the model. In conclusion, WHARMPA might encode global site-effects quantitatively and allow the harmonization of data. This method is user-friendly and provides a powerful solution, without complex implementations, to clean the analyses by removing variability associated with the differences between sites.

KEYWORDS

Gaussian process, machine learning, MRI harmonization, Parkinson's disease, predictive probabilities, site-effects, VBM

[13]Department of Neurology, Knappschaftskrankenhaus Bottrop, Bottrop, Germany

[14]Department of Biomedicine, University of Barcelona, Barcelona, Catalonia, Spain

[15]Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Catalonia, Spain

Correspondence
Medical Psychology Unit, Department of Medicine, University of Barcelona, Casanova 143, 08036, Barcelona, Spain.
Email: bsegura@ub.edu

Funding information
Regarding the University of Barcelona, this study was sponsored by the Spanish Ministry of Economy and Competitiveness (PSI2010-161, PSI2013-41393, PSI2017-86930-P) cofinanced by Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF), and PID2020-114640GB-I00/AEI/10.13039/501100011033, by Generalitat de Catalunya (2017SGR748), Fundació La Marató de TV3 20142310, and supported by María de Maeztu Unit of Excellence (Institute of Neurosciences, University of Barcelona) MDM-2017-0729, Ministry of Science, Innovation and Universities. Regarding the University of Cologne, the study was partially sponsored by the German Research Foundation (project numbers 101434521 and 431549029). Regarding the University of Deusto, this study was sponsored by the Department of Health of the Basque Government (2011111117), the Spanish Ministry of Economy and Competitiveness (PSI2012-32441), and by the A-grade Research Team IT 946-16 from the Basque Government. AS is a consultant for Hoffman La Roche; received honoraria from GE Health Care Canada LTD, Hoffman La Roche. TvE received honoraria by Shire, Lilly, Lundbeck, and Orion Pharma and reports no conflict of interest regarding this study. CU was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie fellowship [grant agreement 888692].

1 | INTRODUCTION

Neuroimaging collections of multi-site Magnetic resonance imaging (MRI) scans are a natural step towards better and deeper knowledge about brain processes in health and disease. Originally, functional, structural or diffusion neuroimaging studies have included small samples, but it has notable limitations. It prevents the detection of true differences, as well as it favors an increased detection of false-positive because of the use of liberal thresholds (Radua et al., 2012). In this sense, large image datasets can be crucial to provide reliable findings, but neuroimaging studies must face the batch effect to harmonize data (Leek et al., 2010). Reproducibility and reliability in multi-site studies are goals that need to be addressed when working with large collections of MRI brain images, mainly to deal with the variance introduced by differences in scanners and acquisition protocols. As a matter of fact, site effects can stunt the detection of consistent findings to the detriment of spurious results (Pinto et al., 2020).

Generally, large-scale datasets are gathered from multiple sites, using different scanners for the acquisition. In many cases, the acquisition protocol is harmonized, and the issue in the final statistical findings seems more restrained (Fox et al., 2012; Vollmar et al., 2010). However, other studies show that diffusion tensor images (DTIs) using

the same acquisition protocol can differ significantly from one site to another (Jovicich et al., 2014; Nyholm et al., 2013). The major drawback, with and without harmonization, is how to remove the non-biological information associated with scanner effects, acquisition protocol and hardware. Beyond biological variability, many relevant works have dealt with scanner manufacturer, scanner upgrade, and field strength (Han et al., 2006), gradient nonlinearity (Jovicich et al. 2006), and longitudinal drift (Takao, Hayashi, & Ohtomo, 2011). These properties of MRI scanners have been explored to reduce bias and variance of brain images (Jovicich et al., 2016). It has been investigated in cortical thickness (Fortin et al., 2018); Han et al., 2006), in voxel-based morphometry studies (Takao, Hayashi, & Ohtomo, 2014), and brain diffusion data (Fortin et al., 2017; Pinto et al., 2020), which has been the modality more studied with many methods. However, the intervention over the scanner and acquisition parameters cannot always protect against the total amount of unwanted variability. In many cases, according to the increased interest for large-scale studies and new collaborations, multi-site data were pooled and integrated after the scanner, and the acquisition parameters were considered together, and data acquired.

Other strategies focusing on pre-processed data have been considered, but these methods work using metrics (volumes, cortical thickness and FA among others) extracted from pre-processed data. In this context, batch-effect correction methods from other areas have been adapted to the metrics from neuroimaging. The ComBat (Johnson, Li, & Rabinovic, 2007) is a batch effect correction tool used in genomics that has been adapted to metrics from diffusion data (Fortin, Parker, et al., 2017), cortical thickness measures (Fortin et al., 2018) and functional connectivity matrices (Yu et al., 2018). The adjusted residual harmonization model is a conventional procedure that relies on using a linear regression model adjusted for biological covariates. ComBat extends this method by modeling site-specific scaling factors and then uses empirical Bayes to improve the stability of site parameters (https://github.com/Jfortin1/ComBatHarmonization). This technique has also been adapted for meta-analyses (Radua et al., 2020) on cortical thickness, surface area and subcortical volumes from the ENIGMA Consortium (http://enigma.ini.usc.edu). Other authors have also extended the ComBat functionality by fitting a generalized additive model (Combat-GAM; Pomponio et al., 2020). Going further, to avoid depending on a statistically representative sample, the Neuroharmony method relies on training a machine learning tool on the ComBat outcomes to capture the relationship between the intrinsic characteristics of the data (Garcia-Dias et al., 2020). Nevertheless, a study by Chen et al. has shown that site differences could remain in covariance patterns after the harmonization, and propose an approach that combines the ComBat method with variance decomposition (CovBat; Chen et al., 2022).

The problem remains opened when the studies aim to explore pre-processed 3D (NIFTI) images instead of metrics. An approach to address harmonization in treated images is the removal of artificial voxel effect by linear (RAVEL) regression method, which is applied after intensity normalization, prior to segmentation (Fortin, Sweeney, Muschelli, Crainiceanu, & Shinohara, 2016). This method is according to two batch effect correction tools used in genomics (SVA and RUV) (Gagnon-Bartsch & Speed, 2012; Leek & Storey, 2007, 2008), and decomposes the voxel intensity into a biological component and an unwanted variation component from a control region from the cerebrospinal fluid. However, the strong dependency on the reference region can be wrongly associated, and the correction might remove biological signals of interest (Fortin et al., 2016).

A widespread approach to deal with multi-site variability in pre-processed images lies in including a confounding variable in the statistical design. This method relies on the same principle in which ComBat does, since it is an extension of the linear regression model adjusted for biological covariates. If assumed, as reported in Chen et al. that ComBat cannot completely remove covariance patterns after its application on the metrics, the same problem could interfere in volumetric studies (Chen et al. 2022). In fact, it has been reported that the performance boost by using the confound versus the models that do not account for the site is poor (Rao, Monteiro, & Mourao-Miranda, 2017). This aspect involves inherently that nonbiological confounders might be interfering in the outcomes, and that the behavior of the site effects may be unpredictable and heterogeneous.

Pattern recognition and machine learning techniques applied to images of any acquisition modality can predict the behavior of a system and identify any encoded pattern of variation associated with the data (Ashburner & Klöppel, 2011). The scope of application for machine learning becomes maximized in the field of deep learning. Sophisticated approaches using deep learning have been developed to face the harmonization challenge in MRI images. Supervised deep image synthesis has been used to create harmonized images, using U-Net neural network architecture to generate synthetic images from an overlapping cohort (DeepHarmony; Dewey et al., 2019). Another proposed generative model to harmonize MRI is cycleGAN-based model (Zhu et al., 2017). However, a limitation of these methods is that they need additional replicated data to train for each site. Also, a deep learning-based training scheme has been used to create scanner-invariant features, by considering harmonization to be multiple sources while maintaining performance on the main task of interest, this procedure reduces the influence of the site on network predictions (Dinsdale et al., 2021). Nevertheless, this framework has some logistical limitations if there is no overlap between datasets, as well as it cannot be easily used in conjunction with tools such as Freesurfer or techniques like the voxel-based morphometry (VBM). These complex approaches are difficult to be applied and adapted to clinical studies, where the main goal is cleaning data efficiently from the site effects without losing biological information. A balance must be accomplished to make the harmonization procedure accessible as well as powerful, without complex implementations.

In this work, we aim to address this problem using a machine learning approach to harmonize structural MRI images, which could also detect covariance patterns associated with the site effects. For the study, we have used a multi-site dataset that involves anonymized MRI data of patients with PD and healthy subjects (HS) from four relevant institutions in PD research. Gaussian process (GP) for machine learning (Rasmussen & Williams, 2006) were used to quantify the site effect for each image. We hypothesized that the outcomes provided

by the GP classification, known as predictive probabilities, encode the parameters that weight the contribution (in terms of similarity) of each site to each image better than the Boolean site covariate of belonging or not to each site, which is 1 (belonging) or 0 (not belonging). These outcomes, namely Weighted HARMonization PArameters (WHARMPA), can then be introduced in the statistical design, which would be equivalent to modifying the Boolean-type covariate by a weight that would encode to what extent, what is common in all the images of a dataset is present in a particular image.

2 | METHODS

The design of the current work involves two main parts. First, the estimation of the harmonization parameters by means of a machine learning classification of the sites. Second, an evaluation of the scope of these parameters to correct (a) a machine learning classification between patients and HS and (b) the comparison of patients versus HS in a voxel-based morphometry study. As references, the same analyses were conducted using also the conventional Boolean site covariate and without involving information about site belonging.

2.1 | Multi-site dataset

For this study, we used a multi-site data set from four centers. Anonymized T1-weighted MRI data of Parkinson's disease (PD) patients and HS were collected from four relevant institutions in PD research: The University of Deusto, Bilbao, Spain (site 1), the University of Barcelona, Barcelona, Spain (site 2), the Center of Addiction and Mental Health (CAMH), Toronto, Canada (site 3), and the University of Cologne, Cologne, Germany (site 4).

From the initial sample, 10 subjects were excluded at the visual inspection: 6 PD and 1 HS due to movement artifacts, and 1 PD and 2 HS due to incomplete acquisition. The final cohort comprised 216 PD patients and 87 HS. The demographic variables considered involved age and sex.

2.2 | Neuroimaging data

The characteristics of the acquisition parameters for the MRI images are described below.

2.2.1 | Dataset from [Site 1]

An MRI scanner Philips Achieva 3 T TX was used for the acquisitions. T1-weighted images were obtained in a sagittal orientation. Repetition time (TR) = 7.4 ms, echo time (TE) = 3.4 ms, matrix size 228 × 218 mm^2; flip angle 9°, field of view (FOV) = 250 mm, slice thickness 1.1 mm, acquisition time = 4′55″, 300 slices, voxel size 0.98 × 0.98 × 0.60 mm^3.

2.2.2 | Dataset from [Site 2]

Acquisitions were made in an 8-channels head coil SIEMENS MAGNETOM TrioTim syngo MR B19 3 T scanner (Siemens). High-resolution three-dimensional (3D) T1-weighted images were acquired a sagittal orientation. TR = 2,300 ms, TE = 2.98 ms, matrix size = 256 × 256 mm^2, flip angle 9°, FOV = 256 mm, acquisition time = 7′48″, 240 slices, voxel size 1.0 × 1.0 × 1.0 mm^3.

2.2.3 | Dataset from [Site 3]

Images were acquired in a General Electrics Discovery MR750 3 T scanner. fast-spoiled gradient echo pulse sequence, in a sagittal orientation. TR = 6.7 ms, TE = 3.0 ms, matrix size 256 × 256 mm^2, flip angle 8°, FOV = 230 mm, acquisition time = 4′16″, 200 slices, voxel size 0.89 × 0.89 × 0.9 mm^3.

2.2.4 | Dataset from [Site 4]

Acquisitions were made in a PRISMA MAGNETOM 3 T scanner (Siemens). Acquisition parameters for T1-weighted structural images, in a sagittal orientation, were as follows: TR = 2,300 ms, TE = 2.32 ms, matrix size 256 × 256 mm^2, flip angle = 8°, FOV = 230 mm, acquisition time = 5′30″, 192 slices, voxel size 0.9 × 0.9 × 0.9 mm^3.

2.3 | MRI preprocessing

For the preprocessing, T1-weighted images were first visually inspected for artifacts and centered in the anterior commissure. Next, the images were segmented into grey matter (GM), white matter (WM), and other tissues, using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). GM and WM data were normalized to the MNI space according to DARTEL technique (Ashburner, 2007), also implemented in SPM12. Finally, Jacobian and non-Jacobian scaled GM and WM images were smoothed using an isotropic Gaussian kernel, using 11 × 11 × 11 mm^3 of FWHM. This size of FWHM was selected based on previous works about the comparison of MRI features (Monte-Rubio et al., 2018).

2.4 | Machine learning classification

2.4.1 | Processing

The Pattern Recognition for Neuroimaging Toolbox (PRoNTo; http://www.mlnl.cs.ucl.ac.uk/pronto/) was used for machine learning analyses (Schrouff et al., 2013).

A multi-class classification approach was applied to perform the site classification. For the subsequent test, the classification between HS

and PD patients, a binary classification was conducted. In both cases, the multiclass and the binary classification problems, we applied the GP machine learning technique (Rasmussen & Williams, 2006).

The outcomes by the GP site classification, known as predictive probabilities, were used as WHARMPA in the subsequent test analyses.

As features, non-Jacobian scaled GM and GM + WM data from the pre-processing were used as spatial patterns in both classification problems. Nonmodulated versions of the tissues have been found to provide better performance in machine learning analyses (Monte-Rubio et al., 2018) than modulated. A single linear kernel was created with the whole dataset, this matrix encodes the dot product between all the images involved in the analysis. As GP is not a multi-kernel approach, when the combination GM + WM was used, respective kernels were concatenated to get a single kernel.

Additionally, the spatial representation of the predictive function (i.e. weight maps), was estimated in all cases. To show the relative contribution of all regions for the model (Schrouff & Mourao-Miranda, 2018), the weight maps were labeled according to the Anatomical Automatic LabellinG (AAL) atlas.

2.4.2 | Statistics

The predicted labels were compared to the true labels using the test dataset, the balanced accuracy (BAC) was obtained to test the performance. The BAC accounts for the number of samples in each class, providing equal weight to the accuracies between classes. Also, class accuracies to show if the model favors some classes over others, and class positive predictive values to represent the number of false positives, are given. Permutation test was run for each classification problem, for demonstration purposes 100 permutations were carried out to associate a p-value = .0099 to the corresponding performance (p-value is equal to $1/R$, it means p-value < .01) (http://www.mlnl.cs.ucl.ac.uk/pronto/prt_manual.pdf). Regarding the PD patients versus HS classification, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve was additionally reported.

Performances from GP classifications were assessed using a leave-one-out cross-validation (LOO-CV) scheme, therefore the whole sample but one was used during the training phase. For each input, a probabilistic label of belonging to one class or another was predicted providing four vectors, one per center. Predictive probabilities from the classification of centers using GM as a feature were used as WHARMPA to the subsequent clinical analyses.

The site information is a categorical variable, and it cannot be introduced into the design as an ordinal vector. It was implemented as a one-hot encoding type, where each value of a variable ("site1", "site2"…) was considered a Boolean 1 (belonging)/0 (not belonging) variable, a column per site is required. In this way, all values were treated with the same importance.

In all the classification analyses, covariates were regressed out, mean-centered and kernel matrices normalized. Regarding the multi-site classification problem, age, sex, and diagnostic group were included as covariates. On the other hand, for the diagnostic classification of PD patients versus HS, covariates involved were sex, age and (a) the WHARMPA, (b) the Boolean site covariate, or (c) no variable accounting for the site.

Regarding the estimation of the performance in the PD versus HS classification, it has been shown that the BAC might be biased in multi-site studies (Solanes et al., 2021), and that machine-learning studies must control for the site-effects after obtaining the BAC provided by the classification. To complete the BAC estimation, the BAC from the site-effect was also estimated using the "multisite.accuracy" R package (https://cran.r-project.org/web/packages/multisite.accuracy). The real target and the predicted target were used to obtain an independent BAC corrected for site-effect, as well the sensitivity and the specificity of the performance.

Finally, McNemar's test was used to provide a p-value for the comparison between harmonization methods in the HS versus PD classification. It was applied to each pair of predicted targets. The McNemar's test is a nonparametric test for paired nominal data. The basis of the test relies on finding out where these two performances differ, comparing the sensitivity and specificity of the two tests on the same sample. The cutoff of significance (α) was set to .05, and the confidence interval (CI) to 95% (SPSS Statistics, V25.0.0.1; IBM Corporation, Armonk, NY).

2.5 | Voxel-based morphometry

2.5.1 | Processing

To explore the scope of the WHARMPA correction in a univariate approach, Voxel-Based Morphometry (VBM; Ashburner & Friston, 2000) analyses were conducted between whole samples of HS and PD. The comparison, using age and sex as covariates, was conducted to obtain the performance according to each type of site-effect correction. Jacobian scaled GM data, smoothed at 11 mm of FWHM, were used for the analyses.

2.5.2 | Statistics

For the VBM analysis, age, sex, and total intracranial volume (TIV) were used as covariates at all analyses. According to the aim of the test, the following were added to the design: (a) the WHARMPA from the GM tissue, (b) the Boolean site covariate or (c) no covariate.

Statistical parametric maps were corrected for multiple comparisons, with a FWE corrected p-value < .05 and an extent threshold of $k = 100$. Tables were elaborated using the xjView visualization tool (https://www.alivelearn.net/xjview).

TABLE 1 Demographics of the whole multi-site sample: Healthy subjects (HS) and Parkinson's disease (PD) patients

	Site 1	Site 2	Site 3	Site 4	Total	ANOVA
HS						
N (f)	26 (11)	29 (15)	17 (8)	15 (9)	87 (43)	
Age ± *SD*	68.31 ± 7.52	64.21 ± 8.77	61.47 ± 7.40	68.40 ± 7.63	65.62 ± 8.29	0.020*
PD						
N (f)	36 (14)	86 (36)	35 (8)	59 (19)	216 (77)	
Age ± *SD*	68.00 ± 6.26	63.94 ± 10.37	64.77 ± 6.41	67.27 ± 10.32	65.66 ± 9.33	0.063
HS versus PD						
Sex (χ^2 *p*-value)	.787	.355	.076	.047*	.027*	
Age (Student's *p*-value)	.861	.902	.104	.693	.971	

*p<.05.

TABLE 2 Scores from center classification

	Multiclass Gaussian Process classification								
Feature set	Balanced accuracy (BAC %)	Class accuracies (%)				Class positive predictive value (%)			
GM	98.39	98.39	96.52	100	98.65	98.39	99.11	100	94.81
WM	95.98	98.39	94.78	96.15	94.59	89.71	99.09	96.15	95.89
GM + WM	98.82	98.39	98.26	100	98.65	98.39	99.12	100	97.33

Note: Permutation test *p*-value = .0099.
Abbreviations: GM, grey matter; WM, white matter.
Feature sets from non-Jacobian scaled data; smooth 11mm FWHM.

3 | RESULTS

3.1 | Demographic data

No significant differences were found regarding age between HS and PD patients (Student's *t*-test, age: T = −0.036, p = .971). Regarding sex, differences were slightly biased between groups (HS: 50.57% females, PD: 64.35% females; Chi-Squared, X^2 test value = 4.922, p = .027; for Site 4: p = .047; Table 1).

3.2 | Estimation of the weighted WHARMPA

The classification of sites provided high scores. Outcomes from each feature are detailed in Table 2.

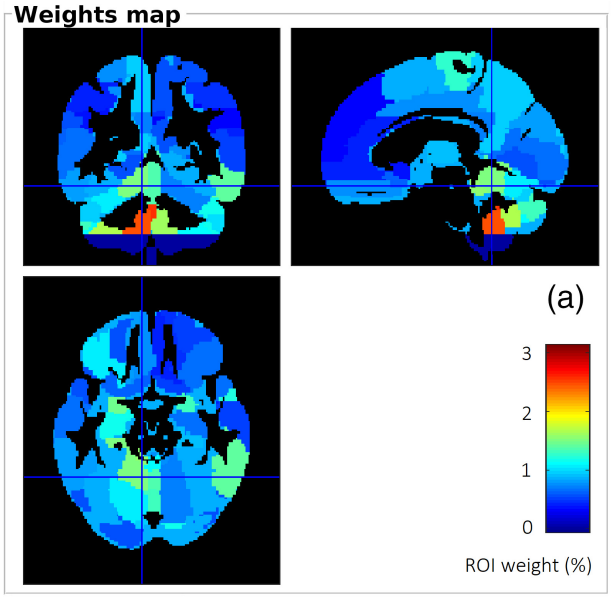
In all cases, the confusion matrix showed a reduced number of classification errors. The class accuracies were different across sites. However, all the features performed very efficiently. The BAC scores were comprised between 95.98 and 98.82%.

Pearson's correlation parameters were estimated between scores from each feature according to sites. A strong correlation between predictive probabilities was found. Lower scores were observed between GM and WM (from Site 1:0.97; Site 2:0.96; Site 3:0.95; and Site 4:0.95). Between GM and GM + WM the correlation was higher (from Site 1:1.00; Site 2:0.99; Site 3:1.00; and Site 4:0.99). Regarding the WM and GM + WM, the scores were in the range (from Site 1:0.98; Site 2:0.98; Site 3:0.97; and Site 4:0.97). We can state that predictive probabilities from the classification using any of these features are highly correlated. However, involving GM seems to provide higher concordance between performances (Figure 1).

3.2.1 | Test 1: Parkinson's disease patients versus healthy subjects classification

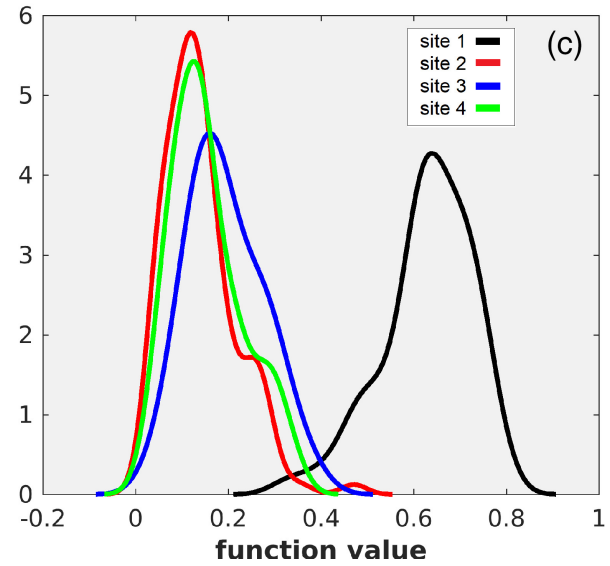
Results from the classification of patients versus controls were quite different depending on the type of site correction (Table 3). Performances from any feature, without correction and using the Boolean site covariate, provided poor efficiency. On the other hand, those analyses involving the WHARMPA showed increased performance. Additionally, when the parameters came from the same feature that the one used in the analysis, results were maximum.

Statistically significant differences were found between the performances from the different harmonization methods according to McNemar's test (Table 4). The Boolean site correction provided better outcomes than when no information about the site was added, although the performance from the WHARMPA method outperformed the rest of the corrections. However, the improvement when using WHARMPA from one tissue or combined was not relevant.



(b)

ROI Label	ROI weight (%)	ROI size (vox)	Exp. Ranking
Vermis_8	3.12	484	117.00
Vermis_9	2.52	335	115.99
Cerebelum_9_L	2.42	922	115.01
Cerebelum_10_L	1.90	115	114.00
Cerebelum_10_R	1.73	108	112.98
Cerebelum_8_L	1.64	1706	111.98
Cerebelum_9_R	1.56	953	110.87
Cerebelum_3_L	1.53	198	109.89
Vermis_7	1.50	415	108.67
Cerebelum_4_5_L	1.49	2233	108.08
ParaHippocampal_L	1.47	1812	106.90
Temporal_Pole_Mid_L	1.47	1376	106.46
Amygdala_L	1.44	432	105.10
Temporal_Inf_R	1.38	7231	103.93
Vermis_4_5	1.34	892	102.80
Vermis_10	1.33	54	101.96
Cerebelum_Crus2_L	1.31	3676	100.66
Vermis_6	1.30	559	100.13



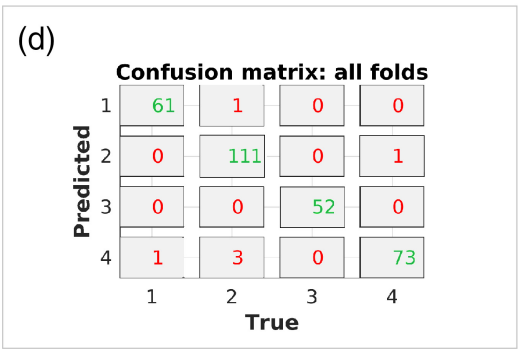


FIGURE 1 (a) Weights map from the site classification using grey matter (GM) as feature; (b) The list shows those regions with a contribution above 100 in the expected ranking (ER) to identify site effects for the current sample; (c) Histogram and (d) confusion matrix from the same classification analysis

The most contributive regions involved in the decision criteria for the GM were mostly in the left hemisphere, the cerebellum 10 (ROI weight 2.10%; size of the ROI 115 voxels; expected ranking [ER] 117) and 9 (ROI weight 2.06%; size of the ROI 922 voxels; ER 115), the left cerebellum Crus2 (ROI weight 1.79%; size of the ROI 3676 voxels; ER 114), the paracentral lobule (ROI weight 2.08%; size of the ROI 1750 voxels; ER 116), also in the right hemisphere (ROI weight 1.47%; size of the ROI 1155 voxels; ER 110). The vermis 1_2 (ROI weight 1.70%; size of the ROI 92 voxels; ER 113) and 9 (ROI weight 1.67%; size of the ROI 335 voxels; ER 112) were also strongly contributing to the decision. A table with the region's relevance with a contribution above 100 in the expected ranking is shown in Figure 2d.

3.2.2 | Test 2: Voxel-based morphometry comparison between PD and HS

In all analyses, only the contrast HS > PD provided results. The outcomes when using the harmonization parameters to correct the model involved more regions, and these were wider than when the Boolean site covariate or no site information were applied. On the contrary, when the Boolean site correction was used, the clusters appeared more spatially restrained and less statistically significant than in the other two models. When no site correction was used, the extent of the regions was sized between both types of harmonizations, and some overlap of clusters with these other two models was observed.

TABLE 3 Scores from the classification between Parkinson's disease (PD) patients and healthy subjects (HS)

Scores derived from	PRoNTo					Multisite accuracy		
Feature for disease classification	Correction for site-effect	Balanced accuracy (BAC %)	Class accuracies (%)	Class predictive value (%)	AUC	BAC (%) EoS corr.	Sens.	Spec.
GM	None	57.22	21.84 92.59	54.29 74.63	0.73	57.21	0.22	0.92
GM + WM	None	58.59	26.44 90.74	53.49 75.38	0.73	58.59	0.26	0.91
GM	Boolean site covariate	56.31	17.24 95.37	60.00 74.10	0.71	56.22	0.17	0.95
GM + WM	Boolean site covariate	56.19	16.09 96.30	63.64 74.02	0.71	57.37	0.18	0.96
GM	WHARMPA from GM	78.60	63.22 93.98	80.88 86.38	0.90	79.09	0.64	0.94
GM + WM	WHARMPA from GM	76.07	58.62 93.52	78.46 84.87	0.90	76.67	0.60	0.93
GM	WHARMPA from GM + WM	76.29	60.92 91.67	74.65 85.34	0.90	77.33	0.63	0.92
GM + WM	WHARMPA from GM + WM	76.30	58.62 93.98	79.69 84.94	0.90	76.73	0.60	0.93

Abbreviations: Boolean site covariate, conventional center information; EoS, effect of site; GM, grey matter; None, no center correction used; Sens, sensitivity; Spec, specificity; WHARMPA, weighted harmonization parameters; WM, white matter.

The VBM analysis between whole samples of HS and PD showed areas bilaterally distributed when the WHARMPA were used to correct for the site effect, also the clusters covered larger regions than when using the conventional center covariate or none. Statistically significant regions appeared more restrained and lateralized when the Boolean site covariate or no site correction were applied.

The statistical map obtained in the VBM using the WHARMPA (colored in green in Figure 3) showed decreased GM volume regions consistent with PD deterioration pattern. Occipital regions, the left occipital middle and superior, and the precuneus and cuneus were comprised in the most significant cluster. The larger cluster was centered in the right frontal superior medial and extended all the frontal lobe and the anterior cingulate. Additionally, the limbic lobe, involving the bilateral cingulum middle towards the supplementary motor area was detected. A large cluster placed in the left frontal inferior orbital that extended towards the temporal superior covered a wide area of the left temporal and part of the left frontal lobe. Moreover, a cluster in the right temporal lobe, involving the middle and superior temporal, extended toward the occipital middle, involving the angular and the cuneus. Predominantly left, but also right thalamus was observed in a cluster that extended toward the hippocampus and then the cerebellum. Bilateral amygdala was also involved.

The VBM analysis without using any type of site correction provided regions (colored in red in Figure 3) that overlapped some of the findings from the VBM using the WHARMPA. The most remarkable location above mentioned in the left occipital regions, as well as the larger cluster in the frontal lobe, were observed. However, the extension of these clusters was rather limited in contrast with the clusters obtained using the WHARMPA. Smaller clusters were also detected in the bilateral temporal superior, and the limbic lobe.

The most relevant difference between the outcomes was observed when comparing the VBM maps using the Boolean site covariate (colored in blue in Figure 3) to the other design models. Only five small clusters were found, the occipital regions were not detected. The larger cluster, placed in the right frontal superior medial, exhibited few voxels with respect to the clusters described above, involving only the superior and medial frontal gyrus. Also, the cluster in the left temporal superior, in the limbic lobe, and in the left temporal inferior involving the fusiform were detected. Statistics and locations involved in the three design models are detailed in Table S5.

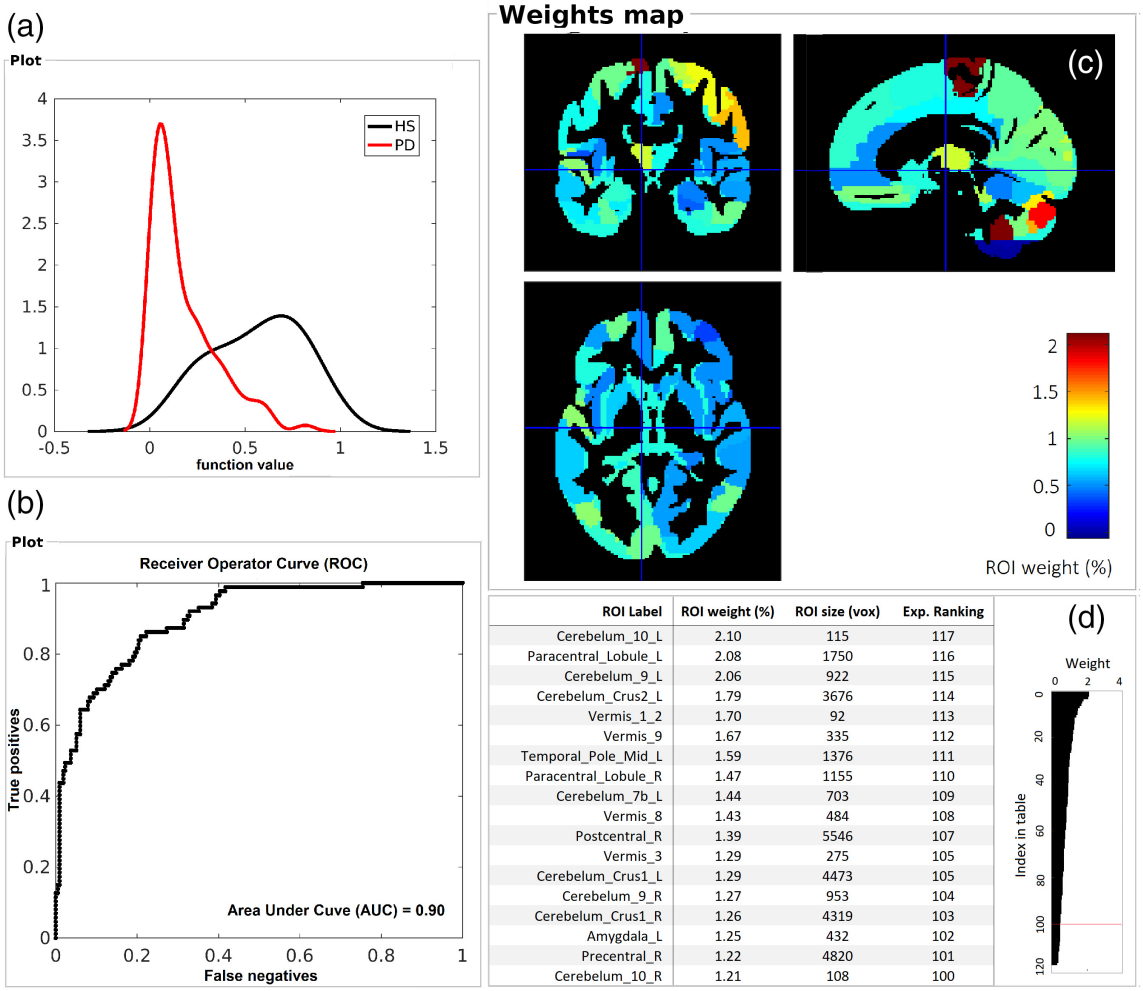
4 | DISCUSSION

In the current work, we have tested the outcomes derived from GP classification as harmonization parameters to correct for the differences between sites. We have observed that these parameters might encode quantitatively the whole variability associated with the site effect, which means those variations associated with the differences in the MRI data. GM and WM tissue images were used as features for the site-effect classification. As expected, both features and its combination provided very similar and high performance, because the target (i.e. site belonging) is a determinant factor. The predictive probabilities

TABLE 4 McNemar's test estimation between each pair of performances

Correction for site-effect		Boolean site covariate		WHARMPA from GM		WHARMPA from GM + WM	
	Feature set	GM	GM + WM	GM	GM + WM	GM	GM + WM
None	GM	0.031*	0.011*	0.000*	0.000*	0.000*	0.000*
	GM + WM	0.002*	0.000*	0.001*	0.002*	0.000*	0.004*
Boolean site covariate	GM		0.581	0.000*	0.000*	0.000*	0.000*
	GM + WM			0.000*	0.000*	0.000*	0.000*
WHARMPA from GM	GM				0.549	0.508	0.388
	GM + WM					0.238	1.000
WHARMPA from GM + WM	GM						0.092
	GM + WM						

Note: Cut-off of significance (α) was set to 0.05, and the confidence interval (CI) to 95%.
Abbreviations: GM, grey matter; None, no site information used; WHARMPA, weighted harmonization parameters; WM, white matter.
*p <.05.



ROI Label	ROI weight (%)	ROI size (vox)	Exp. Ranking
Cerebelum_10_L	2.10	115	117
Paracentral_Lobule_L	2.08	1750	116
Cerebelum_9_L	2.06	922	115
Cerebelum_Crus2_L	1.79	3676	114
Vermis_1_2	1.70	92	113
Vermis_9	1.67	335	112
Temporal_Pole_Mid_L	1.59	1376	111
Paracentral_Lobule_R	1.47	1155	110
Cerebelum_7b_L	1.44	703	109
Vermis_8	1.43	484	108
Postcentral_R	1.39	5546	107
Vermis_3	1.29	275	105
Cerebelum_Crus1_L	1.29	4473	105
Cerebelum_9_R	1.27	953	104
Cerebelum_Crus1_R	1.26	4319	103
Amygdala_L	1.25	432	102
Precentral_R	1.22	4820	101
Cerebelum_10_R	1.21	108	100

FIGURE 2 Scores of the classification between Parkinson's disease (PD) patients and healthy subjects (HS) using grey matter (GM) tissue as feature, and harmonization parameters from GM to protect from the scanner effect: (a) histogram of the classification, (b) ROC and AUC, (c) weights map shows the contribution of each region in the decision and (d) table of the regions that played a role in the classification decision according to the AAL atlas (https://www.pmod.com/files/download/v35/doc/pneuro/6750.htm). An expected ranking (ER) is assigned to each region, in the table only the contributions above 100 were selected

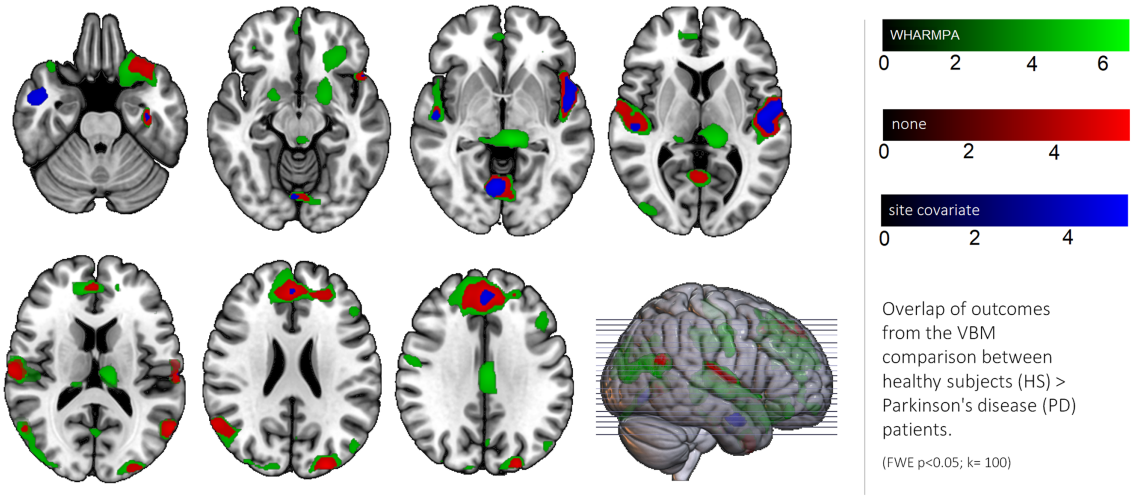


FIGURE 3 Overlap of the outcomes from the comparison between healthy subjects (HS) > Parkinson's disease (PD) patients. Age and sex were used as covariates in all the models. The *t*-score is represented in green for the weighted harmonization parameters (WHARMPA) from grey matter, in blue for the conventional Boolean site correction, and in red when no correction was applied. Right is left

from GM images were used to explore their scope as WHARMPA. The results showed that these encode site-effect variability more accurately than using the conventional Boolean site correction. We tested our hypothesis using a PD multi-site dataset. First, we attempted to identify a classification pattern between patients and controls. Second, we focused on the VBM comparison to identify GM atrophy patterns in the disease. In both cases, when the WHARMPA were used to correct the two different types of clinical analyses, the performance became better than using the Boolean approach.

Regarding the classification between HS and PD patients, the results showed how performance markedly increased when the WHARPMA were used as a correction factor, against the Boolean site regressors and no site correction. The best performance achieved without using the WHARMPA gave a BAC of 58.59% and the AUC was 0.73, which is a poor outcome. However, when the WHARMPA were used to correct the model, the BAC was 79.09% and the AUC of 0.90, this means a statistically significant improvement (McNemar's test, p-value = .001). To put into contrast these scores, there are some studies in the literature that have attempted the classification between patients with PD and HS using whole brain images. In a whole brain resting state fMRI study using between-network connectivity, the authors reported an accuracy of 76.2% (Rubbert et al., 2019). However, DTI data have been found not feasible in a support vector machine study on a whole-brain level (FA, MD), ROI-labeled analyses, or when focusing on the substantia nigra (Prasuhn, Heldmann, Münte, & Brüggemann, 2020). Certainly, the scores can improve when one or few regions obtained in a VBM are later asked to the machine learning analysis. Using this ROI approach, accuracies can reach 99% to detect PD versus HS in men (Solana-Lavalle & Rosas-Romero, 2021). Many studies show increased performances when using non-MRI data, for instance, voice features provided an accuracy of 93.84% (Karapinar Senturk, 2020), or postural sway classification reached 86% (Apthorp et al., 2020). The highest performance that we have obtained using only whole brain data were obtained with the WHARMPA method (BAC 79.09; AUC 0.90) and provided comparable findings from studies that used additional strategies. McNemar's test showed that the proposed method to harmonize studies provided statistically significant improvements with respect to the Boolean regressors or the absence of correction.

Moreover, we found that if the features used to determine the WHARMPA are the same to be used for the posterior study, the coefficients will better fit the data and optimize the performance. Even so, we observed that the predictive probabilities from different features were strongly correlated. In fact, McNemar's test did not show significant differences between performances when using WHARMPA from any feature.

A second test focused on the VBM technique. We showed how harmonizing with WHARMPA provided the best performance concerning amount, size and statistical robustness of the clusters in regions that have been previously reported in PD (Burton, McKeith, Burn, Williams, & O'Brien, 2004; Potgieser et al., 2014). Regions showed a matching with the areas identified in a meta-analysis by Xu et al. (2020). This study found only reduced GM volume in PD patients to HS, no GM volume increase was detected. Authors gathered the regions in five areas, all of them have been observed in the current VBM analyses, but mostly involved in the VBM harmonized using the WHARMPA. These regions involve the bilateral insula, lenticular nucleus and putamen, temporal lobe, right striatum and amygdala. Also, the anterior cingulate, paracingulate gyri, and superior frontal and fusiform gyrus have been reported (Xu et al., 2020). Additionally, the occipital, the bilateral thalamus, basal ganglia and the orbital regions were implicated.

Findings, when no correction was added to the model, were spatially encompassed between those from using the WHARMPA and the Boolean regressors, which were markedly scarce. However, it was expected that the model without site correction provided at maximum the same outcomes as the model with the Boolean site correction.

The lower statistical scores using the Boolean site-regressors may imply that those could be strongly removing nonimaging related (e.g. biological) differences between sites. In this sample, we did not find one variable to be responsible for the nonexpected behavior of the outcomes between the types of correction in our data, some variables might be contributing to this fact.

The influence of unbalanced variables on the harmonization becomes a limitation in these approaches that account for site-effect after preprocessing. This is a general problem related to any multicentric study itself, similar to what happens with cultural, demographic, or genetic differences across countries. The pure modeling of the variance due to the scanner and acquisition protocol is the ideal scenario, but this is quite difficult to reach because often, not all the variables of the multisite dataset are available. The effect becomes wider when comparing the outcomes from using Boolean regressors versus without using site correction. Mostly, without site correction, the results might be strongly biased. In terms of reliability, if something is hidden among the noise, we must struggle with data until we can get the most but, when no real outcomes overcome the threshold, no spurious outcomes must appear. Positively, and even though the WARMPA might be removing also biological differences, this method is efficient in facing this kind of situation better than the Boolean approach that we have explored in this work. Another limitation of this method is that from WHARMPA is not possible to identify what the source of the variance is, whether it is due to the scanner, the protocol, which parameter, and what the amount of biological component is. Thus, all the variables that want to be preserved in the data must be considered when estimating the WHARMPA parameters to avoid removing them.

Further work is being conducted to extend the WHARMPA to metrics, like cortical thickness, volumes or fractional anisotropy (FA). This approach will enable the comparison between methods like ComBat (Johnson et al., 2007), which has shown high efficiency in harmonizing metrics from MRI and DTIs, better than fixed-effects covariates.

In conclusion, the current work shows that the WHARMPA can provide a good solution to accurately clean the analyses in a quantitative way at the image level, by removing variability associated with the site. It is effective in situations that require regress out scanner effects multi-site or local (like a scanner calibration during the acquisition of a cohort). The application of the WHARMPA correction does not require additional implementations and is straightforward, the approach allows an user-friendly obtention of the WHARMPA with a low computational burden.

ACKNOWLEDGMENTS

We acknowledge the Centres de Recerca de Catalunya (CERCA) Program/Generalitat de Catalunya, the Institute of Neurosciences, and the Institute of Biomedical Research August Pi i Sunyer (IDIBAPS). Antonio Strafella was supported by the Canadian Institutes of Health Research (PJ8 169695), the Krembil-Rossy Chair and R. Thomson Foundation. Thilo van Eimeren received honoraria by Shire, Lilly, Lundbeck, and Orion Pharma. Carme Uribe was supported by the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie fellowship [grant agreement 888692].

CONFLICT OF INTEREST

Antonio Strafella is a consultant for Hoffman La Roche; received honoraria from GE Health Care Canada LTD, Hoffman La Roche. CE received honoraria in the last 12 months for consultancy and lectures from Philyra Inc., Abbvie Inc., and Stadapharm Inc.

AUTHORS CONTRIBUTIONS

Gemma Monte-Rubio and Barbara Segura designed the study and drafted the article. Gemma Monte-Rubio and Carme Uribe analyzed the data, and Antonio Strafella, Thilo van Eimeren, Naroa Ibarretxe-Bilbao, María Díez-Cirarda, Carsten Eggers, Olaia Lucas-Jiménez, Natalia Ojeda, Javier Peña, Marina C. Ruppert, Roser Sala-Llonch, Hendrik Theis, and Carme Junque contributed to the interpretation of the data. Antonio Strafella, Thilo van Eimeren, Naroa Ibarretxe-BilbaoRoser Sala-Llonch and Carme Junque revised the manuscript critically for important intellectual content and approved the final version of the manuscript.

ETHICS STATEMENT

All participating sites received approval from an ethical standards committee prior to study initiation.

PATIENT CONSENT STATEMENT

All participating sites obtained written informed consent for research from all participants in the study.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Gemma C. Monte-Rubio https://orcid.org/0000-0002-3532-2224
Barbara Segura https://orcid.org/0000-0002-9673-5479
Antonio P. Strafella https://orcid.org/0000-0002-7779-0974
Thilo van Eimeren https://orcid.org/0000-0002-6951-2325
Naroa Ibarretxe-Bilbao https://orcid.org/0000-0002-2434-5252
Maria Diez-Cirarda https://orcid.org/0000-0001-6768-189X
Carsten Eggers https://orcid.org/0000-0001-7564-6701
Olaia Lucas-Jiménez https://orcid.org/0000-0002-7715-3764
Natalia Ojeda https://orcid.org/0000-0002-0952-0649
Marina C. Ruppert https://orcid.org/0000-0002-9025-7058
Roser Sala-Llonch https://orcid.org/0000-0003-3576-0475
Carme Uribe https://orcid.org/0000-0002-1415-687X
Carme Junque https://orcid.org/0000-0002-6381-3063

REFERENCES

Apthorp, D., Smith, A., Ilschner, S., Vlieger, R., Das, C., Lueck, C. J., & Looi, J. C. L. (2020). Postural sway correlates with cognition and quality of life in Parkinson's disease. *BMJ Neurology Open*, *2*(2), e000086. https://doi.org/10.1136/bmjno-2020-000086

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007

Ashburner, J., & Friston, K. J. (2000). Voxel-Based Morphometry—The methods. *NeuroImage*, *11*(6 Pt 1), 805–821. https://doi.org/10.1006/nimg.2000.0582

Ashburner, J., Klöppel, S. (2011). Multivariate models of inter-subject anatomical variability. *NeuroImage*, *56*(2), 422–439. https://doi.org/10.1016/j.neuroimage.2010.03.059

Burton, E. J., McKeith, I. G., Burn, D. J., Williams, E. D., & O'Brien, J. T. (2004). Cerebral atrophy in Parkinson's disease with and without dementia: A comparison with Alzheimer's disease, dementia with Lewy bodies and controls. *Brain*, *127*(4), 791–800. https://doi.org/10.1093/brain/awh088

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., & Shou, H. (2022). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping*, 858415. *43*(4), 1179–1195. https://onlinelibrary.wiley.com/doi/10.1002/hbm.25688

Dewey, B. E., Zhao, C., Reinhold, J. C., Caras,s A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., van Zijl, P. C.M., & Prince, J. L. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, *64*, 160–170. https://doi.org/10.1016/j.mri.2019.05.041

Dinsdale, N. K., Bluemke, E., Smith, S. M., Arya, Z., Vidaurre, D., Jenkinson, M., & Namburete, A. I. L. (2021). Learning patterns of the ageing brain in MRI using deep convolutional networks. *NeuroImage*, *224*, 117401. https://doi.org/10.1016/j.neuroimage.2020.117401

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Adams, P., … Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*. *167*, https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., … Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., … Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*(July), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Fortin, J. P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., & Shinohara, R. T. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, *132*, 198–212. https://doi.org/10.1016/J.NEUROIMAGE.2016.02.036

Fox, R. J., Sakaie, K., Lee, J. C., Debbins, J. P., Liu, Y., Arnold, D. L., … Fisher, E. (2012). A validation study of multicenter diffusion tensor imaging: Reliability of fractional anisotropy and diffusivity values. *American Journal of Neuroradiology*, *33*(4), 695–700. https://doi.org/10.3174/ajnr.A2844

Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics (Oxford, England)*, *13*(3), 539. https://doi.org/10.1093/BIOSTATISTICS/KXR034

Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W. H. L., Corvin, A., … Mechelli, A. (2020). Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *NeuroImage*, *220*. 1–15. https://doi.org/10.1016/J.NEUROIMAGE.2020.117127

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., … Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, *32*(1), 180–194. https://doi.org/10.1016/j.neuroimage.2006.02.051

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Jovicich, J., Marizzoni, M., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., … Frisoni, G. B. (2014). Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects. *NeuroImage*, *101*, 390–403. https://doi.org/10.1016/j.neuroimage.2014.06.075

Jovicich, J., Minati, L., Marizzoni, M., Marchitelli, R., Sala-Llonch, R., Bartrés-Faz, D., … Frisoni, G. B. (2016). Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: A multicentric resting-state fMRI study. *NeuroImage*, *124*(Pt A), 442–454. https://doi.org/10.1016/j.neuroimage.2015.07.010

Karapinar Senturk, Z. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. *Medical Hypotheses*, *138*, 109603. https://doi.org/10.1016/j.mehy.2020.109603

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., … Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733–739. https://doi.org/10.1038/nrg2825

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, *3*(9), 1724–1735. https://doi.org/10.1371/JOURNAL.PGEN.0030161

Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(48), 18718–18723. https://doi.org/10.1073/PNAS.0808709105

Nyholm, T., Jonsson, J., Söderström, K., Bergström, P., Carlberg, A., … Zackrisson, B. (2013). Variability in prostate and seminal vesicle delineations defined on magnetic resonance images, a multi-observer, −center and -sequence study. *Radiation Oncology*, *8*(1), 1–12. https://doi.org/10.1186/1748-717X-8-126

Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P. J., Jeurissen, B., … Sijbers, J. (2020). Harmonization of brain diffusion MRI: Concepts and methods. *Frontiers in Neuroscience*, *14*, 396. https://doi.org/10.3389/fnins.2020.00396

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., … Davatzikos, C. (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

Potgieser, A. R. E., van der Hoorn, A., Meppelink, A. M., Teune, L. K., Koerts, J., & de Jong, B. M. (2014). Anterior temporal atrophy and posterior progression in patients with Parkinson's disease. *Neurodegenerative Diseases*, *14*(3), 125–132. https://doi.org/10.1159/000363245

Prasuhn, J., Heldmann, M., Münte, T. F., & Brüggemann, N. (2020). A machine learning-based classification approach on Parkinson's disease diffusion tensor imaging datasets. *Neurological Research and Practice*, *2*(46), 1–5. https://doi.org/10.1186/s42466-020-00092-y

Radua, J., Mataix-Cols, D., Phillips, M. L., El-Hage, W., Kronhaus, D. M., Cardoner, N., & Surguladze, S. (2012). A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European Psychiatry*, *27*(8), 605–611. https://doi.org/10.1016/j.eurpsy.2011.04.001

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., … Pineda-Zapata, J. (2020). Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage*, *218*(May), 116956. https://doi.org/10.1016/j.neuroimage.2020.116956

Rao, A., Monteiro, J. M., & Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, *150*, 23–49. https://doi.org/10.1016/j.neuroimage.2017.01.066

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. London, England: The MIT Press.

Rubbert, C., Mathys, C., Jockwitz, C., Hartmann, C. J., Eickhoff, S. B., Hoffstaedter, F., … Caspers, J. (2019). Machine-learning identifies Parkinson's disease patients based on resting-state between-network

functional connectivity. *The British Journal of Radiology*, *92*(1101), 20180886. https://doi.org/10.1259/bjr.20180886

Schrouff, J. and Mourão-Miranda, J. (2018). Interpreting weight maps in terms of cognitive or clinical neuroscience: nonsense? *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 1–4. https://doi.org/10.1109/PRNI.2018.8423944

Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., … Mourão-Miranda, J. (2013). PRoNTo: Pattern recognition for neuroimaging toolbox. *Neuroinformatics*, *11*(3), 319–337. https://doi.org/10.1007/s12021-013-9178-1

Solana-Lavalle, G., & Rosas-Romero, R. (2021). Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, *198*, 105793. https://doi.org/10.1016/j.cmpb.2020.105793

Solanes, A., Palau, P., Fortea, L., Salvador, R., González-Navarro, L., Llach, C. D., … Radua, J. (2021). Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. *Psychiatry Research: Neuroimaging*, *314*, 111313. https://doi.org/10.1016/J.PSCYCHRESNS.2021.111313

Takao, H., Hayashi, N., & Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. *Journal of Magnetic Resonance Imaging*, *34*(2), 438–444. https://doi.org/10.1002/jmri.22636

Takao, H., Hayashi, N., & Ohtomo, K. (2014). Effects of study design in multi-scanner voxel-based morphometry studies. *NeuroImage*, *84*, 133–140. https://doi.org/10.1016/j.neuroimage.2013.08.046

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., … Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. *NeuroImage*, *51*(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046

Xu, X., Han, Q., Lin, J., Wang, L., Wu, F., & Shang, H. (2020). Grey matter abnormalities in Parkinson's disease: A voxel-wise meta-analysis. *European Journal of Neurology*, *27*(4), 653–659. https://doi.org/10.1111/ene.14132

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., … Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, *39*(11), 4213. https://doi.org/10.1002/HBM.24241

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *IEEE International Conference on Computer Vision (ICCV)*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: C. Monte-Rubio, G., Segura, B., P. Strafella, A., van Eimeren, T., Ibarretxe-Bilbao, N., Diez-Cirarda, M., Eggers, C., Lucas-Jiménez, O., Ojeda, N., Peña, J., Ruppert, M. C., Sala-Llonch, R., Theis, H., Uribe, C., & Junque, C. (2022). Parameters from site classification to harmonize MRI clinical studies: Application to a multi-site Parkinson's disease dataset. *Human Brain Mapping*, *43*(10), 3130–3142. https://doi.org/10.1002/hbm.25838