# Fed-ComBat: A Generalized Federated Framework for Batch Effect Harmonization in Collaborative Studies

Santiago Silva*[a], Ghiles Reguig*[a], Neil P. Oxtoby[b], Andre Altmann[b], Marco Lorenzi[a]

[a] *Université Côte d'Azur and INRIA, EPIONE team, Sophia Antipolis, France*
[b] *University College London, The UCL Hawkes Institute, Department of Medical Physics and Biomedical Engineering, London, UK*

---

**Abstract**

The use of multi-centric analyses is crucial for obtaining sufficient sample sizes and representative clinical populations in experimental studies. In this setting, data harmonization techniques are typically employed to address systematic biases and ensure the interoperability of the data. State-of-the-art harmonisation approaches are based on the statistical theory of random effect modeling, allowing to account for either linear of non-linear biases and batch effects. However, optimizing these statistical methods generally requires data centralization at some point during the analysis pipeline, therefore introducing the risk of exposing individual patient information while posing significant data governance issues. To overcome this challenge, in this paper we present Fed-ComBat, a federated framework for batch effect harmonization on decentralized data. Fed-ComBat enables the preservation of nonlinear covariate effects without requiring centralization of data and without prior parametric hypothesis on the variables to account for. We demonstrate the

---

*Email address:* `marco.lorenzi@inria.fr` (Marco Lorenzi)
*Authors with equal contribution

effectiveness of Fed-ComBat against a comprehensive panel of existing approaches based on the state-of-the-art ComBat, along with distributed and nonlinear variants. Our experiments are based on extensive simulated data, and on the analysis of multiple cohorts based on 7 neuroimaging studies comprising healthy controls (CI) and subjects with various disorders such as Parkinson's disease (PD), Alzheimer's disease (AD), and autism spectrum disorder (ASD).

Our results show that in a federated settings, Fed-ComBat harmonization exhibits comparable results to centralized methods for both linear and nonlinear cases. On real data, harmonized trajectories of the thickness of the right hippocampus across lifespan measured on a set of 7 public studies show comparable results between centralized and federated models and are consistent with the literature when using a nonlinear model.

The code is publicly available at: `https://gitlab.inria.fr/greguig/fedcombat`

*Keywords:* Harmonization, federated learning, medical imaging, multi-centric data, batch effects, data privacy, ComBat

---

## 1. Introduction

Data harmonization is known to be a crucial factor in tackling data heterogeneity in studies across multiple sites (Wachinger et al., 2021). Statistical data harmonization aims to address biases and variations among data collected from different sources, ensuring that the data is comparable and can be combined effectively for analysis. Typical harmonization methods, such as ComBat, aim at correcting potential biases due to site effects while pre-

serving the ones associated with covariates of interest (e.g., sex, diagnosis, age) (Fortin et al., 2017, 2018). The standard formulation of ComBat relies on linear mixed effect models to disentangle the site variability from the fixed effects associated with the desired covariates. More recently, ComBat-GAM extended this framework by reformulating the linear mixed effect functions through generalized additive models (GAMs) to account for nonlinear covariate effects Pomponio et al. (2020). Although ComBat-GAM showed better performance than ComBat in the LIFESPAN dataset (Pomponio et al., 2020), it remains limited by design on the parameterization of the non-linear functions, such as polynomial or splines. Moreover, ComBat-GAM requires an explicit definition of covariate interactions, typically limited to additive or multiplicative terms (e.g., Sex $\times$ Diagnosis). Despite efforts, a gap still exists in the availability of methods for flexible covariate effect preservation without requiring data centralization Gebre et al. (2023).

This problem is further exacerbated by the practical setting of real-world multi-centric studies. Due to governance standards and privacy issues, data often cannot be shared among sites in a centralized server. This issue challenges the practical use of standard harmonisation protocols that generally assume data availability. Federated learning (FL) is a popular machine learning paradigm that allows for collaborative model optimization while maintaining data privacy and governance (Konečný et al., 2016). In FL, models are trained on data that remains decentralized across multiple institutions or devices, ensuring that sensitive data stays securely within the respective entities (Konečný et al., 2016). FL operates by sharing only model updates or parameters, aggregating these updates on a central server, and distributing

the updated global model back to each institution.

d-ComBat is a recent extension of ComBat to adapt the harmonisation task to the federated setting Chen et al. (2022). The underlying idea of this approach is to adapt the optimization of ComBat in the distributed setting through the decomposition of the data covariance matrix across sites. This operation allows solving the least squares problem associated to the linear ComBat function, subsequently allowing the estimation of local sites effects while bypassing the data sharing task. Nevertheless, d-ComBat can be applied only in the linear setting, due to the specific optimization scheme tailored to the ordinary least squares (OLS) problem. The extension of d-ComBat to non-linear effect estimation thus requires the investigation of more general distributed optimization schemes compatible with the ComBat formulation. This is essential to allow the generalization of the harmonisation task to account for non-linear covariate effects in real-world multi-site settings Bethlehem et al. (2022).

To fill this methodological gap, in this work we investigate a general formulation of ComBat based on flexible and non-parametric models of covariate effects, while allowing for a distributed optimization in a federated learning setup. Our approach enables a more nuanced representation of covariate effects during the harmonization process while encompassing ComBat, ComBat-GAM, and d-ComBat as specific cases.

We benchmarked our proposed methods with existing centralized, distributed, linear and non-linear variants (Johnson et al., 2007; Pomponio et al., 2020; Chen et al., 2022). The different methods were compared for their ability to harmonize batch effects and preserve the quality of covariate effects

on simulated data (Section 4). Furthermore, we performed an evaluation on derived phenotypes from MRI-brain images from seven cohorts, comprising controls and patients with different brain disorders: patients with different subtypes of Parkinson's disease (PD), Alzheimer's disease (AD) and Autism spectrum disorder (ASD).

In all our experiments, our proposed models achieved accurate and generally superior harmonization capabilities while maintaining biological covariates information.

Ultimately, our work extends current possibilities to data harmonization in multi-centric clinical studies to real decentralized setups that are compliant with data governance and sharing constraints.

## 2. Methods

### *2.1. Generalized ComBat model*

Following the original ComBat formulation proposed by Johnson et al. (2007), let us denote a batch (represented by different scanners protocols, machines, or institutions) indexed by $i \in \{1, 2, ..., S\}$ on a particular phenotype (e.g., a brain region) indexed by $g \in \{1, 2, ..., G\}$. Each batch contains $n_i$ number of observations, and the total number of observations is $N = \sum_i^S n_i$. $S$ can denote for simplicity the number of sites in the study, but it can also be extended to the total number of scanners between sites or any other number of batch effects. We can model a specific phenotype $g$ observed in the $j$-th patient who belongs to the $i$-th site denoted by $y_{ijg}$ as follows:

$$y_{ijg} = \alpha_g + \phi(\mathbf{x}_{ij}, \boldsymbol{\theta}_g) + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}, \tag{1}$$

where $\mathbf{x}_{ij}$ denotes the covariate effects expected to be preserved after removing the batch effects (e.g., sex and age), $\alpha_g$ acts as a global fixed intercept (i.e., the mean), while $\gamma_{ig}$ indicates a random intercept that accounts for the site-specific shift. $\varepsilon_{ijg}$ is a noise model that captures the variability of each phenotype $\varepsilon_{ijg} \sim \mathcal{N}(0, \sigma_g^2)$, and $\delta_{ig}$ is a multiplicative effect that scales the "unbiased" phenotype variability to fit the one at each site.

This formulation generalizes the original linear model proposed by Johnson et al. (2007) and the nonlinear covariate effect model proposed by Pomponio et al. (2020), to account for potentially multivariate and linear fixed functions $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ parameterized by $\boldsymbol{\theta}_g$.

We introduce the following constraints to the estimation of fixed effect parameters $(\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}, \hat{\sigma}_g)$:

$$\arg \max_{\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}, \hat{\sigma}_g} P(y_{ijg} | \hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}, \hat{\sigma}_g) \tag{2}$$

$$\text{subject to} \quad \mathbb{E}_g[\hat{\gamma}_i] = \sum_i^S \frac{n_i}{N} \hat{\gamma}_{ig} = 0, \, \forall g \in \{1, ..., G\} \tag{3}$$

$$\text{and} \quad \phi(\mathbf{x}, \boldsymbol{\theta}_g)|_{x=0} = 0 \tag{4}$$

A first constraint in Equation (3) is set to allow identifiability of the intercept parameters as explained by Johnson et al. (2007). However, in the same regard, the constraint in Equation (4) ensures that no batch effects are captured by the covariate function $\phi(\cdot)$. This second constraint, despite not being mentioned, is also fulfilled by Johnson et al. (2007) and Pomponio et al. (2020); making ComBat and ComBat-GAM a particular case of the proposed formulation in this work.

For a centralized setup, the estimation of all these parameters is per-

formed in three steps: i) maximum likelihood estimation (MLE) for parameters $\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}$ (see Equation (2)) and for the phenotype variance $\hat{\sigma}_g^2 = \frac{1}{N}\sum_{ij}\left(y_{ijg} - \hat{\alpha}_g - \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g) - \hat{\gamma}_{ig}\right)^2$, ii) residual standardization mapping the residuals to satisfy the form $y_{ijg} \to z_{ijg} \sim \mathcal{N}(\gamma_{ig}, \delta_{ig}^2)$ as follows:

$$z_{ijg} = \frac{y_{ijg} - \hat{\alpha}_g - \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g)}{\hat{\sigma}_g} \tag{5}$$

and iii) estimation of the additive and multiplicative batch effects $\hat{\gamma}_{ig}^*$ and $\hat{\delta}_{ig}^*$ as in Equation (6), using empirical Bayes (EB) with priors on $\gamma_{ig}$ and $\delta_{ig}^2$ to iteratively estimate these parameters as in Equation (7) (Johnson et al., 2007, Sec. 3.2).

$$\gamma_{ig} \sim \mathcal{N}(Y_i, \tau_i^2) \quad \text{and} \quad \delta_{ig}^2 \sim \text{Inverse Gamma}(\lambda_i, \vartheta_i) \tag{6}$$

$$\gamma_{ig}^* = \frac{n_i \bar{\tau}_i^2 \hat{\gamma}_{ig} + \delta_{ig}^{2*} \bar{\gamma}_{ig}}{n_i \bar{\tau}_i^2 + \delta_{ig}^{2*}}, \; \delta_{ig}^* = \frac{\bar{\theta}_i + \frac{1}{2}\sum_{j=1}^{n_i}(z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_i}{2}\bar{\lambda}_i - 1} \tag{7}$$

Lastly, phenotypes can be harmonized while preserving the covariate effects of interest as follows:

$$y_{ijg}^{\text{ComBat}} = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}\left(z_{ijg} - \hat{\gamma}_{ig}^*\right) + \hat{\alpha}_g + \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g) \tag{8}$$

In the following section, we will discuss how this formulation facilitates the incorporation of harmonization within the federated learning framework.

### *2.2. Federated learning*

In a standard federated setting we focus on a distributed optimization problem associated with the functional:

$$\arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}), \tag{9}$$

$$\text{where, } F(\boldsymbol{\theta}) := \sum_{i=1}^{S} \frac{n_i}{N} F_i(\boldsymbol{\theta}), \tag{10}$$

where $F_i$ is the cost function evaluated at site $i$, $n_i$ is the number of data points available at site $i$, and $N = \sum_i n_i$. *Federated averaging* (FEDAVG) is the standard approach to optimize the problem of Equation (10) (McMahan et al., 2017), based on the iteration of partial optimization of problem (10) at each site followed by a step of parameters aggregation across clients. In this setting, when local optimization is performed by stochastic gradient descent (SGD), the convergence of the federated optimization process has been demonstrated for both IID and non-IID data distribution across clients (Li et al., 2019).

### *2.3. SGD-ComBat: Flexible and General ComBat optimization*

We propose the SGD-ComBat model by defining the kernel function $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ as a neural network optimized with stochastic gradient descent (SGD). This allows the harmonization process to model covariate effects (defined by the kernel function) using universal approximators such as multi-layer perceptrons (MLP).

Two variants of SGD-ComBat are used in our work: a linear one, which we refer to as SGD-ComBat Linear, and a non-linear one, modeling the covariate effects with an MLP which we refer to as SGD-ComBat MLP. To respect the constraint mentioned in Equation 4, the layers of the SGD models do not include biases.

Once the SGD model has been optimized, the parameters $\boldsymbol{\theta} = \{\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}\}$ are estimated as described in the Equations 7.

### 2.4. From SGD-ComBat to Fed-ComBat

The extension of SGD-ComBat to the federated learning regime follows directly by replacing the global SGD optimization step with an iterative federated scheme, such as FedAVG or FedProx (Li et al., 2020). In particular, Considering the formulation presented in Equation (1), the parameters $\boldsymbol{\theta} = \{\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}\}$ can be optimized as in Equation (10), followed by the local estimation of the random effects following the standard ComBat routine, for example based on Empirical Bayes (EB) (Equations (6) and (7)). A description of the steps followed in federated harmonization using Fed-ComBat is described in Algorithm 1 including local updates relying on SGD, global updates across sites of shared parameters using FedAVG and random effect estimation.

## 3. Materials

We evaluated SGD-ComBat and Fed-ComBat on a synthetic benchmark accounting for different sources of batch- and covariate-wise heterogeneity as well as different data/covariates generation schemes. In addition we benchmarked this approach on a collection of seven cohorts corresponding to different studies in neurodegenerative disorders including participants diagnosed with or at risk of developing Parkinson's disease, Alzheimer's disease and Autism spectrum disorder.

### 3.1. Synthetic data

Following the previous formulation of Equation (1), data was generated as such:

$$y_{ijg} = \alpha_g + \phi(\mathbf{x}_{ij}, \boldsymbol{\theta}_g) + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

**Algorithm 1:** Fed-ComBat

**Data:** Non-harmonized phenotypes ($y_{ijg}$) and covariates ($\mathbf{x}_{ij}$).

**Result:** Harmonized phenotypes with siloed data.

$\mathbf{x} \leftarrow$ FederatedStandardization($\mathbf{x}$);

Estimation of fixed effects and random intercept:

initialize of parameter space $\boldsymbol{\Omega} := \{\boldsymbol{\theta}_g, \alpha_g, \gamma_{ig}\}_{g=1}^{G}$;

**while** *not converged* **do**

  **foreach** *site i* **do**

    $\boldsymbol{\Omega}_i^t \leftarrow \boldsymbol{\Omega}^{(t+1)}$ ; `// Initialization.`

    `// Partial local optimization using SGD.`

    **foreach** *local gradient step t* **do**

      $\boldsymbol{\Omega}_i^{(t+1)} = \boldsymbol{\Omega}_i^{(t)} - \eta \nabla_{\boldsymbol{\Omega_i}} F(\boldsymbol{\Omega_i}^{(t)})$ ;

  `// Aggregate and update every parameter using FedAVG.`

  $\boldsymbol{\Omega}^{(t+1)} \leftarrow \sum_i \frac{n_i}{N} \boldsymbol{\Omega}_i^t$

**forall** *site i* **do**

  `// Standardize using estimated parameters`

  $z_{ijg} \leftarrow \frac{y_{ijg} - \hat{\alpha}_g - \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g)}{\hat{\sigma}_g}$

  `// Estimate` $\gamma_{ig}^*$ `and` $\delta_{ig}^*$ `using EB`

  $\gamma_{ig}^*, \delta_{ig}^* \leftarrow$ EmpiricalBayes($\mathbf{x}$) ; `// Equations (6) and (7)`

  `// Correct data`

  **return** $y_{ijg}^{\text{ComBat}} \leftarrow \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}\left(z_{ijg} - \hat{\gamma}_{ig}^*\right) + \hat{\alpha}_g + \phi(\mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}_g)$

with

- $\gamma_i \sim \text{Uniform}(-5, 5)$.
- $\delta_{ig} \sim \text{Uniform}(2, 5)$ and $\varepsilon_{ijg} \sim \text{Uniform}(1, 4)$.
- $\phi$ is a kernel function with parameters $\boldsymbol{\theta}_g \sim \text{Uniform}(-10, 10)$ allowing to modulate the dependence of the phenotype $y_{ijg}$ with the covariates $\mathbf{x}_{ij} \sim \mathcal{N}(0; 1)$

### *3.2. Brain MRI-Data*

We evaluated our method on a cross-sectional cohort of 13479 participants from seven public studies, utilizing MRI-derived phenotypes from baseline structural T1-weighted magnetic resonance imaging (MRI).

Detailed demographic information can be found in Table 1. Covariate distribution across the multiple cohorts used in this work can be observed in Figure 1, clearly denoting the non-IID nature of this dataset.

#### *3.2.1. Image processing*

For all datasets, with the exception of ABIDE I, cortical thickness and subcortical volume were extracted from MRI scans using FreeSurfer v7.1.1 (documented and freely available for download online at: `http://surfer.nmr.mgh.harvard.edu/`) (Dale et al., 1999; Fischl et al., 2004). Steps for phenotypical measures extraction included skull stripping (Ashburner and Friston, 2005), Non-uniform intensity Normalization (N3) (Sled et al., 1998), segmentation using the Desikan–Killiany atlas (Desikan et al., 2006) and extraction of the MRI-derived phenotypes.

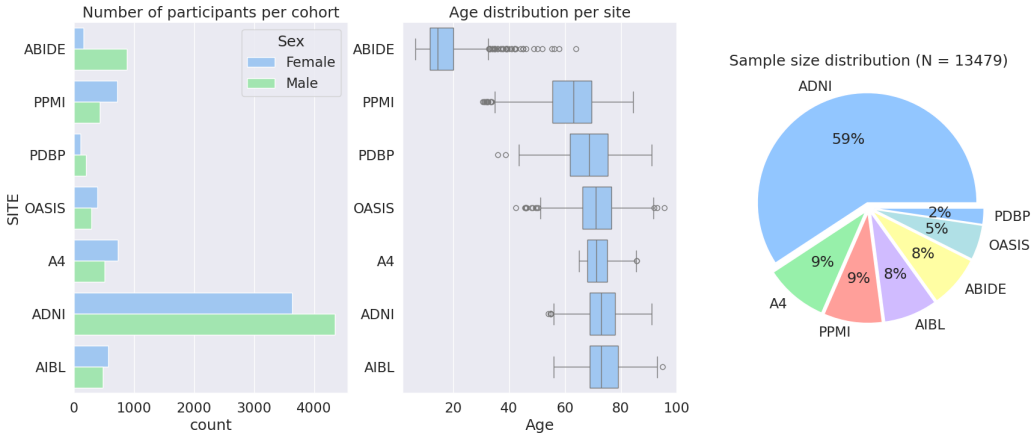


Figure 1: Population pyramid representing the demographics of the real data used in this study. The left panel shows the distribution of sex, the middle panel shows the distribution of age, and the right panel shows the sample size distribution. Cohorts were sorted by ascending median age. The demographics for the population are described in Table 1.

The ABIDE I phenotypes were downloaded from the related sources[2] and were extracted using FreeSurfer v6. We chose to include the ABIDE data in our analysis to experiment with additional batch effects and variability, and to emphasize the non linearity induced by the differences in brain development patterns between young and elder populations.

Harmonization was carried out by preserving age, sex, diagnosis and intracranial volume (ICV / eTIV), and each study was considered a site whose systematic biases were expected to be corrected (Reynolds et al., 2022).

[2] `https://github.com/dfsp-spirit/abide_preproc_smri_freesurfer6`

Table 1: Subject demographics of real data used in this work.

| | | N | | Age in years ± SD [age range] | |
|---|---|---|---|---|---|
| **Site** | **Group \ Sex** | **Female** | **Male** | **Female** | **Male** |
| **A4** | **CU** | 731 | 515 | 71.3 ± 4.5 [65.0 - 85.7] | 72.9 ± 5.1 [65.0 - 85.7] |
| **ABIDE** | **Autism** | 62 | 443 | 16.3 ± 7.9 [8.1 - 45.0] | 17.2 ± 8.6 [7.0 - 64.0] |
| | **CU** | 95 | 435 | 14.8 ± 5.5 [7.8 - 32.0] | 17.2 ± 7.8 [6.5 - 56.0] |
| **ADNI** | **AD** | 710 | 891 | 73.1 ± 7.8 [55.0 - 91.0] | 74.5 ± 7.0 [55.0 - 90.0] |
| | **CU** | 1477 | 1327 | 72.8 ± 6.0 [55.0 - 90.0] | 73.9 ± 6.2 [60.0 - 90.0] |
| | **MCI** | 1451 | 2128 | 71.5 ± 7.7 [55.0 - 88.0] | 73.6 ± 7.0 [54.0 - 90.0] |
| **AIBL** | **AD** | 60 | 55 | 75.4 ± 7.4 [58.0 - 93.0] | 76.0 ± 7.6 [60.0 - 89.0] |
| | **CU** | 445 | 345 | 72.8 ± 6.5 [60.0 - 91.0] | 73.6 ± 6.6 [60.0 - 92.0] |
| | **MCI** | 67 | 85 | 76.0 ± 7.8 [56.0 - 95.0] | 75.0 ± 6.0 [64.0 - 88.0] |
| **OASIS** | **AD** | 59 | 76 | 76.1 ± 6.9 [62.9 - 95.6] | 76.5 ± 6.9 [62.9 - 95.6] |
| | **CU** | 333 | 218 | 68.8 ± 8.7 [45.6 - 93.1] | 70.7 ± 8.3 [42.5 - 91.9] |
| **PDPB** | **PD** | 75 | 120 | 69.4 ± 8.2 [50.8 - 90.0] | 68.1 ± 10.6 [38.8 - 91.0] |
| **PPMI** | **CU** | 34 | 42 | 66.0 ± 10.1 [50.7 - 84.4] | 64.7 ± 11.2 [36.0 - 84.3] |
| | **LBD** | 4 | 38 | 74.0 ± 6.9 [67.9 - 81.9] | 68.1 ± 8.0 [45.1 - 82.7] |
| | **GenCohort PD** | 45 | 38 | 65.4 ± 8.5 [32.2 - 78.6] | 65.5 ± 8.0 [51.7 - 81.2] |
| | **GenCohort Unaff** | 47 | 73 | 61.1 ± 7.9 [48.8 - 78.3] | 61.2 ± 8.4 [33.7 - 84.3] |
| | **PD** | 415 | 207 | 62.5 ± 9.8 [34.8 - 84.2] | 61.3 ± 9.5 [33.5 - 81.7] |
| | **Prodromal** | 32 | 5 | 69.7 ± 6.3 [60.8 - 84.0] | 70.7 ± 2.7 [67.1 - 73.2] |
| | **SWEDD** | 54 | 28 | 63.6 ± 9.4 [39.3 - 78.9] | 60.3 ± 10.1 [39.5 - 79.6] |
| | **CU** | 130 | 84 | 60.7 ± 11.8 [30.6 - 82.8] | 58.9 ± 10.6 [31.0 - 81.9] |

## 4. Results

Two versions of Fed-ComBat were used: the first one sets the kernel as a linear model (Fed-ComBat Linear) and the second one sets $\phi$ as a multi-layer perceptron (Fed-ComBat MLP). These two models were compared with their centralized versions optimized with gradient descent, which we refer to as SGD-ComBat Linear for the linear model and SGD-ComBat MLP for the

| Model | Non Linear | Distributed |
|---|---|---|
| NeuroComBat | ✗ | ✗ |
| SGD-ComBat Linear | ✗ | ✗ |
| d-ComBat | ✗ | ✓ |
| Fed-ComBat Linear | ✗ | ✓ |
| ComBat-GAM | ✓ | ✗ |
| SGD-ComBat MLP | ✓ | ✗ |
| Fed-ComBat MLP | ✓ | ✓ |

Table 2: Methods used in our experiments and their characteristics

non linear one.

ComBat[3] (a.k.a. NeuroComBat), ComBat-GAM[4] as well as the distributed version d-ComBat[5] were used for comparison. The characteristics of the models are reported in Table 2. All implementations were done in Python.

### *4.1. Synthetic data*

In this experiment, the covariates $\mathbf{x}_{ij}$ were simulated in two dimensions and three strategies were used in the generation of the unbiased phenotype $\phi(\mathbf{x}_{ij}, \boldsymbol{\theta}_g)$:

- linear generation: the kernel is a simple linear model

$$\phi_{linear}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{x}^\intercal \boldsymbol{\theta}$$

.

[3] `https://github.com/Jfortin1/neuroCombat`
[4] `https://github.com/rpomponio/neuroHarmonize`
[5] `https://github.com/andy1764/Distributed-ComBat`

- nonlinear generation: the kernel is a quadratic model

$$\phi_{quadratic}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{x}^\intercal\boldsymbol{\theta} + (\mathbf{x}\odot\mathbf{x})^\intercal\boldsymbol{\theta}$$

  where $\odot$ is the Hadamard product.

- mixed generation: the kernel is a mix of a linear model and a quadratic model

$$\phi_{mixed}(\mathbf{x};\boldsymbol{\theta}) = \phi_{linear}(x_1;\boldsymbol{\theta_1}) + \phi_{quadratic}(x_2;\boldsymbol{\theta_2})$$

The number of clients/sites was set to 10 and the simulation was repeated 50 times. The MLP is composed of two hidden layers of 50 neurons, a first one followed by a ReLU activation and a second one followed by a TanH, and a final output layer. The learning rate was set at $10^{-3}$ with 300 epochs for the centralized models (SGD-ComBat Linear and SGD-ComBat MLP) and 100 rounds of 3 epochs for the federated models.

The residuals were computed as

$$residuals = y_{ijg}^{\text{ComBat}} - \phi(\mathbf{x}_{ij};\hat{\boldsymbol{\theta}}_g) - \hat{\alpha}_g = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}\big(z_{ijg} - \hat{\gamma}_{ig}^*\big).$$

The standard deviations of the residuals of each of the 50 runs were computed for each model. Their distributions are reported on Figure 2 for linear, non linear and mixed simulated data. We observe, as expected, that linear models (in blue) show higher standard deviation residuals in the presence of non linear data whether it is from the non linear or mixed simulation. Regarding the non linear models, ComBat-GAM yields the lowest standard deviations of the residuals and both our proposed non linear methods (SGD ComBat MLP and Fed-ComBat MLP) obtain similar results which are really

close to the ComBat-GAM ones.

Figure 3 shows the Bland-Altmann plot for a given non-linear simulation. The Figure shows a clear distinction of data heterogeneity between linear and nonlinear models. We also observe that the federated models achieve similar results to the centralized ones both visually and in RMSE values.

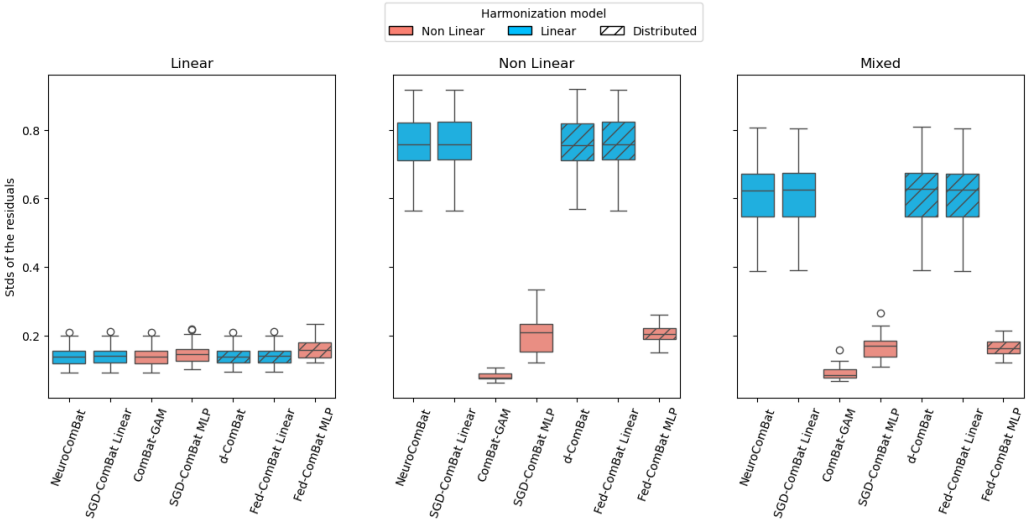


Figure 2: Boxplot of the means of the residuals over 50 runs of simulation. The left, center and right plots respectively show the results for linear, non-linear and mixed simulated data. We note that federated models achieve the similar results as their centralized counterparts, for both linear and nonlinear data.

### *4.2. Brain MRI-data*

To provide evidence for nonlinear covariate effects of age on brain phenotypes, we compared the goodness of fit of two models: a linear model and a generalized additive model (GAM). Data was first centralized and harmonized accounting for study as the batch source using ComBat. The criterion used to evaluate the presence of nonlinearities was the difference

Figure 3: Bland-Altman plots and root mean square errors (RMSE) contrasting the different centralized and federated harmonization methods against the groundtruth. We note that the linear models all show higher RMSE compared to the nonlinear ones. The nonlinear models also show more homogeneous data distributions.

of the Akaike Information Criterion (AIC) between the GAM and the linear model, where a negative difference indicates the improved fit of the non-linear GAM. Figure 5 shows the AIC metric across the regions of the brain using the Desikan-Killiany parcellation and cortical thickness across regions and the ASEG atlas for subcortical volumes as the dependent variables. Data was controlled for sex, diagnosis, ICV, and diagnosis group, and age was considered an independent variable. Figure 5 shows the top four regions that are better explained by a nonlinear model of age in terms of AIC, as well as the residuals after the regression illustrating remaining effects or trends.

Then data was decentralized and we illustrate the age trajectories by regrouping participants in two trajectories: cognitively unimpaired were labeled as “Healthy controls“ while those diagnosed with a particular neuro-

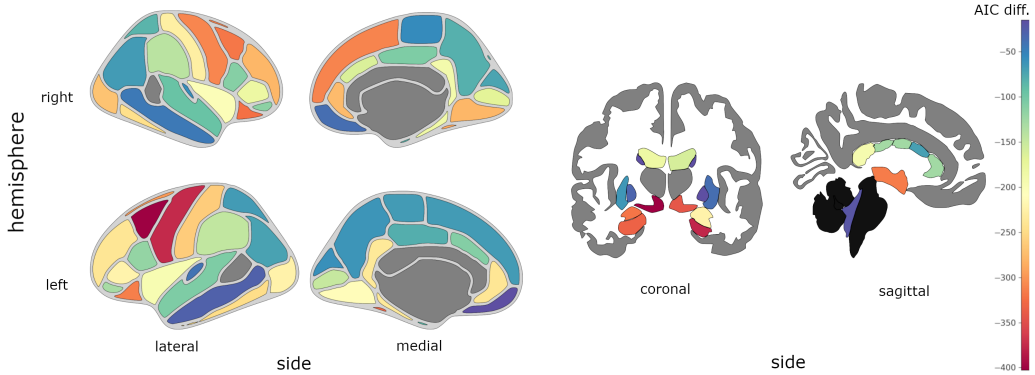


Figure 4: Brain regions where a generalized additive model (GAM) provides a better fit than a linear model, as measured by the difference in Akaike Information Criterion (AIC) between the two models. The parameters for the GAM are set as in Pomponio et al. (2020), with age as a smoothing term, B-splines with 10 degrees of freedom or control points uniformly distributed between the minimum and maximum values, and a maximum polynomial of 3rd degree. A negative difference indicates a better fit with the GAM, while a positive difference indicates a better fit with the linear model.

logical disorder were labeled as "Atypical". The harmonized phenotypes were obtained using NeuroComBat as linear centralized method, ComBat-GAM as nonlinear centralized, d-ComBat as linear federated method, our adaptation Fed-ComBat using a linear model or an MLP, and finally Fed-ComBat using the same MLP architecture as described previously (two hidden layers of 50 neurons followed by a ReLU activation and a final output layer).

After estimation of each harmonisation method, we assessed the resulting age trajectories by fitting a GAM on the harmonised data by controlling for Sex and ICV. Resulting trajectories are illustrated in Figure 6. The results show that all linear models yield similar trajectories with a continuous decrease throughout lifespan, shown by uniform slope in the trajectory. Regarding the nonlinear models, our proposed Fed-ComBat MLP shows a slight decrease in the beginning followed by a steeper drop starting around 40/50, while our centralized proposed model SGD MLP shows a slow increase until 50/60 before dropping. Finally, the ComBat-GAM model shows a steadier curve until 50/60 before decreasing. Even though the nonlinear trajectories seem conflicting, they also reflect the non consensus of the evolution of

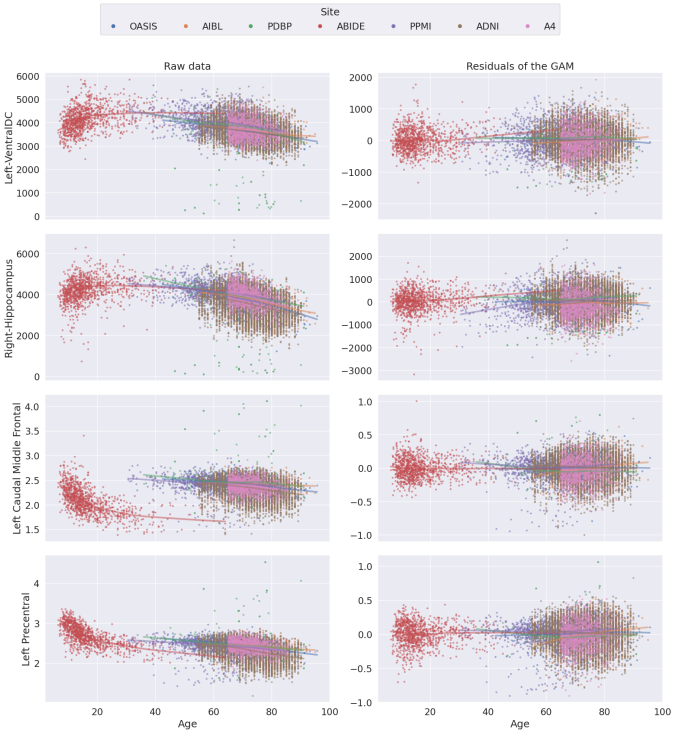


Figure 5: Results of regression using a GAM on the top 4 regions with the most discrepant AIC values on the left and residual plots illustrating goodness of fit of a GAM on the right. From the top to the bottom row: 1) Left ventral diencephalon; 2) Right hippocampus; 3) Left caudal middle frontal gyrus; 4) Left precentral gyrus. On the left, we observe nonlinearity of the trajectories with different rates of atrophy across lifespan. On the right, we see the residuals of the fitted GAM which are centered towards 0 across lifespan, showing that a GAM suitably learns the nonlinearity of the trajectories

cortical thickness present in the litterature where some works show a slow decrease since early age (Ducharme et al., 2016) and others show an inverted U-shape evolution (Coupé et al., 2017).

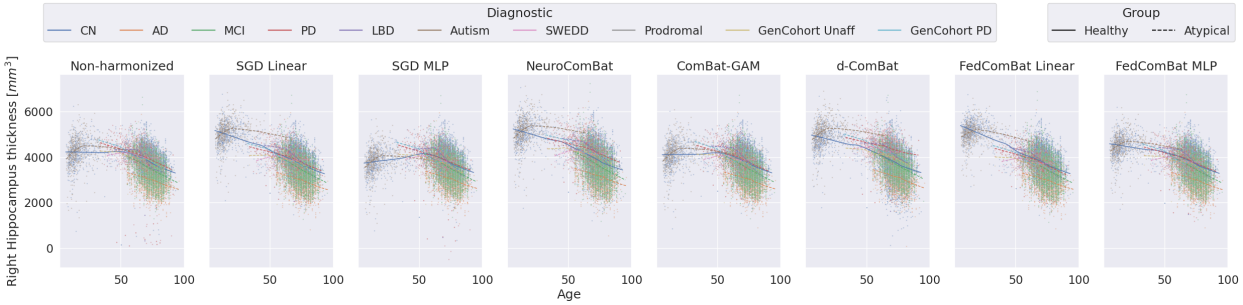


Figure 6: Comparison of age trajectories of the right hippocampus. The columns represent conditions under which the data is shown: non-harmonized (first column), and harmonized using SGD-ComBat Linear (centralized linear SGD based model), SGD-ComBat MLP (centralized using MLP), NeuroComBat, ComBat-GAM, d-ComBat, Fed-ComBat Linear (federated equiv. to d-ComBat but SGD based) and Fed-ComBat MLP (decentralized). Trajectories were adjusted for sex and ICV. We observe that harmonization using linear models (SGD Linear, NeuroComBat and Fed-ComBat Linear) show atrophy patterns that are not consistent with the literature while the nonlinear ones (SGD MLP, ComBat-GAM and Fed-ComBat MLP) better match the different rates of atrophy across lifespan regarding controls (CN).

Figure 7 shows the residuals of each distributed model comprehended in this work after harmonization. The residual harmonized term $\frac{\hat{\sigma}_g}{\hat{\delta}^*_{ig}}\left(z_{ijg} - \hat{\gamma}^*_{ig}\right)$ should follow a normal standard noise distribution by definition and should not have center residual effects. Pomponio et al. (2020) proposed fitting a GAM and then extract the residuals. However, by definition, fitting a GAM will work better with a ComBat model that is GAM-based. Instead, we use the corresponding model for each method (e.g., linear for ComBat, GAM

for ComBat-GAM and MLP for Fed-ComBat MLP). While for each method the distribution of the residuals is centered around 0, the residual variance across dataset is significantly reduced for Fed-ComBat and ComBat-GAM as compared to NeuroComBat ($p < 0.05$ Bartlett test).

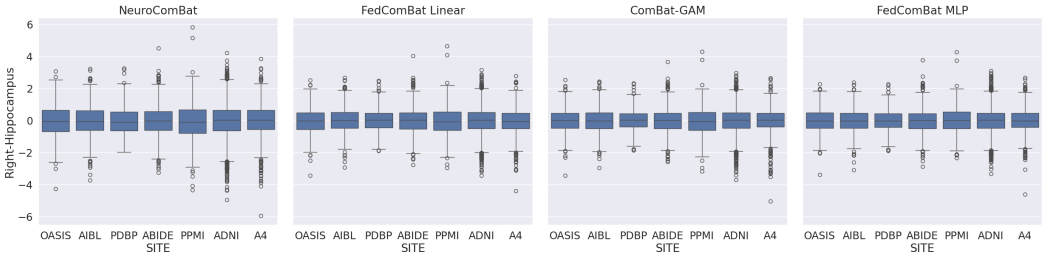


Figure 7: Residual terms after harmonization of the thickness of the right hippocampus for federated methods and their centralized equivalent from the literature. While for each method the distribution of the residuals is centered around 0, the residual variance across dataset is significantly reduced for Fed-ComBat and ComBat-GAM as compared to NeuroComBat ($p < 0.05$ Bartlett test).

## 5. Discussion and Conclusion

In this paper, we presented Fed-ComBat, a generalized approach for data harmonization using federated learning. Specifically designed for scenarios where data sharing is restricted and nonlinear covariate effects might be present, Fed-ComBat provides an extension to existing centralized linear methods such as NeuroComBat. It incorporates the advantages of these methods without necessitating data centralization or explicit specification of nonlinearities and potential interactions.

The performance of Fed-ComBat, both in terms of its linear and multi-layer perceptron (MLP) versions, is comparable to that of centralized meth-

ods. This was evident from the similarity in Bland-Altman plots, despite subtle differences due to the inherent variability in federated setups compared to centralized ones. The MLP version of Fed-ComBat efficiently captured nonlinearities, offering a more generalizable solution than ComBat-GAM.

Moreover, Fed-ComBat does not require explicit decomposition of covariates or specification of interactions between them, unlike ComBat-GAM. This flexibility is facilitated by the capacity of complex approximating functions, such as MLPs, to capture nonlinearities and interactions due to its multivariate nature. The general formulation of the ComBat method that we proposed allows the use of a wider variety of models to better suit the specificities of each harmonization problem, as we have shown using an MLP in our experiments.

While this work showed the effectiveness of Fed-ComBat in handling complex harmonization problems, future research directions could involve evaluating Fed-ComBat's performance on larger cohorts, such as in Bethlehem et al. (2022). This would provide additional evidence of Fed-ComBat's effectiveness in addressing harmonization problems, particularly in scenarios where nonlinear effects are prevalent as was shown for the evolution of several brain regions throughout the entire lifespan by Coupé et al. (2017).

## 6. Limitations and Future Directions

Fed-ComBat has shown to be a promising tool for data harmonization in federated learning settings, but it has some inherent limitations that suggest avenues for future exploration. Besides the exploration of new or improved models, required work to make this method available in platforms like

COINSTAC (Plis et al., 2016) and Fed-BioMed (Silva et al., 2020) will allow real-scenario applications where our approach could be properly evaluated.

Another aspect to consider is the heterogeneity across participating institutions, which could impact the performance of Fed-ComBat, and of ComBat-based correction methods in general. Extreme disparities in data distributions, demographic compositions, or data collection protocols might challenge the method's ability to rectify biases efficiently. This indicates the need to appraise the applicability of Fed-ComBat or any ComBat method in such situations carefully. Furthermore, the current implementation employs Stochastic Gradient Descent (SGD) for optimization which can be improved, especially when dealing with such heterogeneous data. To overcome this, future work could explore methods like FedProx (Li et al., 2020) that provide an efficient and robust solution to SGD in the presence of data heterogeneity.

Beyond these limitations inherent to Fed-ComBat, there are challenges associated with federated learning in general, such as the requirement for significant computational resources and stable connectivity at each site. Acknowledging these, future research could focus on enhancing the robustness of Fed-ComBat against such practical constraints.

The current implementation of Fed-ComBat focuses on derived phenotypes, mirroring the first version of ComBat. However, it would be worthwhile to investigate the method's direct application to imaging data in the future. By modifying $\phi(\mathbf{x}_{ij}, \boldsymbol{\theta}_g)$ to a convolutional neural network (CNN), Fed-ComBat could be adapted to handle image data directly, which would create new possibilities for harmonization in multi-centric imaging studies in federated setups. These considerations will potentially enhance the utility of

Fed-ComBat, extending its reach in federated learning and data harmonization within healthcare research.

Simultaneously to this work, the recent work of Gebre et al. (2023) illustrates the difficulties that harmonization methods face when dealing with imaging data, especially longitudinal data with scanner variations. Their study found that statistical methods like ComBat could match the distributions of cortical thickness across scanners but could not increase intra-class correlation coefficients (ICCs) or eliminate scanner change effects in longitudinal data sets. Deep learning methods like CycleGAN and Neural Style Transfer (NST) performed slightly better, but no method could completely harmonize the longitudinal dataset, underlining the complexity and difficulty of this problem. This suggests the necessity for more sophisticated methods capable of addressing these challenges centralized and inherently, federated scenarios. A recent deep learning-based ComBat harmonization method (DeepComBat) was proposed by Hu et al. (2024), using a conditional variational autoencoder (CVAE) to model both nonlinear and multivariate relation between covariates and cortical thickness measurements in a centralzed setting. Our proposed federated framework provides a generic optimization scheme that can be adapted in the future to extend this methodology to a decentralized setting.

## 7. Ethics statement

This study analyzes publicly available datasets. All data collection protocols from the original studies were approved by their respective institutional review boards (IRB).

## 8. Data and Code Availabilty

The simulation code and the implementation of the proposed algorithms are available at: `https://gitlab.inria.fr/greguig/fedcombat`.

## 9. Author Contributions

S.S., G.R. and M.L. implemented the methods and conducted the experiments, while N.P.O. and A.A. contributed to the preprocessing of the data. S.S., N.P.O., A.A. and M.L. contributed to devising the core idea of the approach. N.P.O., A.A., and M.L. contributed to the revision of the manuscript. M.L. and A.A. contributed to the funding of the project. All authors read and approved the manuscript.

## 10. Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 847579 (Marie Skłodowska-Curie Actions), and by the INRIA Sophia Antipolis-Méditerranée "NEF" computation cluster. This work was partially supported by the Early Detection of Alzheimer's Disease Subtypes (E-DADS) project, an EU Joint Programme - Neurodegenerative Disease Research (JPND) project (see www.jpnd.eu). The project is supported under the aegis of JPND through the following funding organizations: United Kingdom, Medical Research Council (MR/T046422/1); Netherlands, ZonMW (733051106); France, Agence Nationale de la Recherche (ANR-19-JPW2-000); Italy, Italian Ministry of Health (MoH); Australia, National Health & Medical Research Coun-

cil (1191535); Hungary, National Research, Development and Innovation Office (2019-2.1.7-ERA-NET-2020-00008).

## 11. Declaration of Competing Interests

The authors declare no competing interests.

## 12. Acknowledgments

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`https://adni.loni.usc.edu/`). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at `https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; BristolMyers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F.

Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (`https://adni.loni.usc.edu/aibl-australian-imaging-biomarkers-and-lifestyle-study-of-ageing-18-month-data-now-released/`). Data used in the preparation of this article were obtained from the Autism Brain Imaging Data Exchange (ABIDE) I database. The AIBL researchers contributed data but did not participate in the analysis or writing of this report. AIBL researchers are listed at `https://www.aibl.csiro.au`. Data used in the preparation of this article were obtained from the Open Access Series of Imaging Studies (OASIS) database. Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (`http://www.ppmi-info.org/`). The A4 Study was a secondary prevention trial in preclinical Alzheimer's disease, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 Study was funded by a public-private-philanthropic partnership, including funding from the National Institutes of

Health-National Institute on Aging, Eli Lilly and Company, Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation, and additional private donors, with in-kind support from Avid Radiopharmaceuticals, Cogstate, Albert Einstein College of Medicine and the Foundation for Neurologic Diseases. The companion observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) Study was funded by the Alzheimer's Association and GHR Foundation. The A4 and LEARN Studies were led by Dr. Reisa Sperling at Brigham and Women's Hospital, Harvard Medical School, and Dr. Paul Aisen at the Alzheimer's Therapeutic Research Institute (ATRI) at the University of Southern California. The A4 and LEARN Studies were coordinated by ATRI at the University of Southern California, and the data are made available under the auspices of Alzheimer's Clinical Trial Consortium through the Global Research & Imaging Platform (GRIP). The complete A4 Study Team list is available on: `https://www.actcinfo.org/a4-study-team-lists/`. We would like to acknowledge the dedication of the study participants and their study partners who made the A4 and LEARN Studies possible.

## References

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Bethlehem, R.A., Seidlitz, J., White, S.R., Vogel, J.W., Anderson, K.M., Adamson, C., Adler, S., Alexopoulos, G.S., Anagnostou, E., Areces-Gonzalez, A., et al., 2022. Brain charts for the human lifespan. Nature 604, 525–533.

Chen, A.A., Luo, C., Chen, Y., Shinohara, R.T., Shou, H., Initiative, A.D.N., et al., 2022. Privacy-preserving harmonization via distributed combat. Neuroimage 248, 118822.

Coupé, P., Catheline, G., Lanuza, E., Manjón, J.V., Initiative, A.D.N., 2017. Towards a unified analysis of brain maturation and aging across the entire lifespan: A mri analysis. Human brain mapping 38, 5501–5518.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage 9, 179–194.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31, 968–980.

Ducharme, S., Albaugh, M.D., Nguyen, T.V., Hudziak, J.J., Mateos-Pérez, J.M., Labbe, A., Evans, A.C., Karama, S., Group, B.D.C., et al., 2016. Trajectories of cortical thickness maturation in normal brain development—the importance of quality control procedures. Neuroimage 125, 267–279.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., et al., 2004. Automatically parcellating the human cerebral cortex. Cerebral cortex 14, 11–22.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al., 2018. Har-

monization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.

Fortin, J.P., Parker, D., Tunç, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., et al., 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170.

Gebre, R.K., Senjem, M.L., Raghavan, S., Schwarz, C.G., Gunter, J.L., Hofrenning, E.I., Reid, R.I., Kantarci, K., Graff-Radford, J., Knopman, D.S., et al., 2023. Cross–scanner harmonization methods for structural mri may need further work: A comparison study. NeuroImage 269, 119912.

Hu, F., Lucas, A., Chen, A.A., Coleman, K., Horng, H., Ng, R.W., Tustison, N.J., Davis, K.A., Shou, H., Li, M., et al., 2024. Deepcombat: A statistically motivated, hyperparameter-robust, deep learning approach to harmonization of neuroimaging data. Human brain mapping 45, e26708.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8, 118–127.

Konečnỳ, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D., 2016. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 .

Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V., 2020. Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems 2, 429–450.

Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z., 2019. On the convergence of fedavg on non-iid data. arXiv preprint arXiv:1907.02189 .

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR. pp. 1273–1282.

Plis, S.M., Sarwate, A.D., Wood, D., Dieringer, C., Landis, D., Reed, C., Panta, S.R., Turner, J.A., Shoemaker, J.M., Carter, K.W., et al., 2016. Coinstac: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. Frontiers in neuroscience 10, 365.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., Fan, Y., et al., 2020. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage 208, 116450.

Reynolds, M., Chaudhary, T., Torbati, M.E., Tudorascu, D.L., Batmanghelich, K., 2022. Combat harmonization: Empirical bayes versus fully bayes approaches. bioRxiv .

Silva, S., Altmann, A., Gutman, B., Lorenzi, M., 2020. Fed-biomed: A general open-source frontend framework for federated learning in healthcare, in: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI

2020, Lima, Peru, October 4–8, 2020, Proceedings 2, Springer. pp. 201–210.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE transactions on medical imaging 17, 87–97.

Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A.D.N., et al., 2021. Detect and correct bias in multi-site neuroimaging datasets. Medical Image Analysis 67, 101879.