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

Santiago Silva[a], Neil P. Oxtoby[c], Andre Altmann[b], Marco Lorenzi[a,], for the Alzheimer's Disease Neuroimaging Initiative

[a] *Université Côte d'Azur and INRIA, EPIONE team, Sophia Antipolis, France*
[b] *Centre for Medical Image Computing (CMIC), Department of Medical Physics and Biomedical Engineering, University College London, London, UK*
[c] *UCL Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK*

Abstract

In neuroimaging research, the use of multi-centric analyses is crucial for obtaining sufficient sample sizes and representative clinical populations. Data harmonization techniques are typically employed in the pipeline of multi-centric studies to address systematic biases and ensure the interoperability of the data. However, most multi-centric studies require data centralization at some point during the analysis pipeline. Therefore, introducing a risk of exposing individual patient information. This poses a significant challenge in data governance. To mitigate the risk of exposing patient information, various privacy regulations have been introduced, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Although these regulations address such governance concerns, they also hinder researchers from accessing data. Federated learning (FL) offers a privacy-aware alternative approach for machine learning which enables mod-

Email address: `marco.lorenzi@inria.fr` (Marco Lorenzi)

els to be collaboratively trained on decentralized data without the need for data centralization. Although FL provides a solution to the issue of data centralization, it is still crucial to ensure data harmonization for FL models to work effectively with decentralized data.

In this paper, we present Fed-ComBat, a federated framework for batch effect harmonization on decentralized data. Fed-ComBat extends existing centralized linear methods, such as ComBat, distributed like d-ComBat, and nonlinear approaches like ComBat-GAM in accounting for potentially nonlinear and multivariate covariate effects. By doing so, Fed-ComBat enables the preservation of nonlinear covariate effects without requiring centralization of data and without prior knowledge of which variables should be considered nonlinear or their interactions, differentiating it from ComBat-GAM. We assessed Fed-ComBat and existing approaches on simulated data and multiple cohorts comprising healthy controls (CN) and subjects with various disorders such as Parkinson's disease (PD), Alzheimer's disease (AD), and autism spectrum disorder (ASD).

The results indicate that our nonlinear version of Fed-ComBat outperforms centralized ComBat in the presence of nonlinear effects and is comparable to centralized methods such as ComBat-GAM. Using synthetic data, Fed-ComBat can better reconstruct the target unbiased function by 35% (RMSE = 0.5952) with respect to d-ComBat (RMSE = 0.9162) and 12% with respect to our proposed federated d-ComBat-GAM (RMSE= 0.6751) and exhibits comparable results on MRI-derived phenotypes to centralized methods as ComBat-GAM without the need of prior knowledge on potential nonlinearities.

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

1. Introduction

Neuroimaging studies, especially those incorporating machine learning (ML) techniques, greatly benefit from large sample sizes. However, the current approach of centralizing data for analysis faces significant challenges due to data protection regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) (European Commission, 2016; Bukaty, 2019; Calzada, 2022). Despite efforts in anonymization, such as defacing, it is important to note that these techniques may introduce alterations in certain image-derived phenotypes, and some report re-identification rates of around 30%. (Schwarz et al., 2021). This underscores the need for a reevaluation of the traditional paradigms for multi-centric clinical studies, which heavily rely on data centralization.

Data harmonization is known to be a crucial factor in tackling certain challenges of multi-centric data analysis due to systematic errors or *batch effects* commonly present in such studies (Wachinger et al., 2021). Harmonization aims to address biases and variations among data collected from different sources (e.g., scanner brands or institutions), ensuring that the data is comparable and can be combined effectively for analysis. Johnson et al. (2007) proposed ComBat in the field of gene expression analysis, as a way to correct these batch effects while preserving the desired covariate effects (e.g., sex, diagnosis, age). Later, this method was adapted to brain magnetic resonance imaging (MRI) phenotypes (Fortin et al., 2017, 2018). However,

the typical use of ComBat in multi-centric studies still requires data to be centralized, thus limiting the application to real-life collaborative analysis scenarios. To overcome this problem, Chen et al. (2022) proposed d-ComBat to harmonize data within a distributed setup by extending the optimization of ComBat to allow bias correction without the need to exchange the data across centers. To achieve this, d-ComBat shares full local covariance matrices computed at each site, and the subsequently aggregates these covariance matrices to estimate the effect parameters at a global level. This method was shown to achieve almost an exact solution with respect to its centralized version requiring only a few rounds of communication. However, sharing full covariance matrices represents a potential risk of sensitive data leakage (van Kesteren et al., 2019), and like ComBat, d-ComBat considers that covariates have a linear influence on phenotypes, which may not always be the case as reported by Bethlehem et al. (2022) where age shows a clear nonlinear influence on MRI-derived phenotypes. To handle nonlinear covariate effects, Pomponio et al. (2020) introduced ComBat-GAM, which relies on generalized additive models (GAMs) for the preservation of such nonlinear effects. By using smoothing functions, such as polynomial or splines, to decompose covariate effects into basis functions, ComBat-GAM approximates the aforementioned covariate influences. Although ComBat-GAM showed better performance than ComBat in the LIFESPAN dataset (Pomponio et al., 2020), it remains limited by design as it has been conceived only for centralized settings. 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 nonlinear covariate effect preservation without requiring data centralization. Furthermore, as highlighted by Gebre et al. (2023) in a study conducted at the Mayo Clinic, the harmonization problem remains unsolved, calling for further research and collaborative efforts in the field to address this challenge.

Federated learning (FL) is a machine learning paradigm that allows for collaborative model optimization while maintaining data privacy and governance (Konečnỳ 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 (Konečnỳ 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. This iterative process continues until convergence is achieved. Federated Averaging (FedAvg) is a popular approach for aggregation in federated learning. It adapts the stochastic gradient descent (SGD) optimization algorithm for the federated setting providing similar convergence guarantees (McMahan et al., 2017). FL facilitates access to larger cohorts and it has successfully been applied in neuroimaging for tumor segmentation and diagnosis (Li et al., 2019a; Mahlool and Abed, 2022), disease studies in functional magnetic resonance imaging (fMRI) (Li et al., 2020b), intracranial hemorrhage (ICH) detection(Cheung et al., 2023) and continues to be increasingly adopted in neuroimaging applications.

An opportunity opens to create adaptable and efficient methods for standardizing neuroimaging data in federated learning (FL) setups, while also responding to the aforementioned challenges in harmonization. We propose

first a generalized formulation of ComBat that extends to multivariate and nonlinear covariate effect modeling by allowing the incorporation of more complex functions. This enables a more nuanced representation of covariate effects during the harmonization process while encompassing ComBat and ComBat-GAM as specific cases. Secondly, we propose Fed-ComBat, a generalized federated ComBat framework designed specifically for data batch effect harmonization in federated settings. This framework enables the harmonization of data while preserving privacy and security within distributed environments, aligning with the principles of federated learning.

We benchmarked and compared our proposed methods with existing centralized approaches (ComBat, and ComBat-GAM) and distributed (d-ComBat) (Johnson et al., 2007; Pomponio et al., 2020; Chen et al., 2022). Moreover, for the sake of comparison we also formulated and implemented a distributed version of ComBat-GAM (d-ComBat-GAM), which relies on distributed covariance estimation, and extend the method of Chen et al. (2022) by allowing basis decomposition to interpolate nonlinear covariate effects using smoothing functions.

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 nine cohorts: A4 (Sperling et al., 2014), ABIDE-I (Di Martino et al., 2014), ABIDE-II (Di Martino et al., 2017), ADNI (Weiner et al., 2013), AIBL (Ellis et al., 2009), MIRIAD (Malone et al., 2013), OASIS3 (LaMontagne et al., 2019), PDPB (Rosenthal et al., 2016), PPMI (Marek et al., 2018). Comprising controls and people with different

brain disorders: patients with different subtypes of Parkinson's disease (PD), Alzheimer's disease (AD) and Autism spectrum disorder (ASD).

Using synthetic data, we find that Fed-ComBat achieves an RMSE of 0.9162(i.e., 35%) improvement in reconstructing the target unbiased function compared to d-ComBat, and a 12% improvement compared to our federated extension of ComBat-GAM, d-ComBat-GAM (RMSE = 0.6751). In addition, our application on real data shows that Fed-ComBat on MRI-derived phenotypes is comparable to centralized ComBat-GAM. This implies that the studied cohorts may not have exhibited complex nonlinear effects or interactions that could not be captured by a GAM. However, Fed-ComBat requires no prior assumptions of non-linearities unlike ComBat-GAM and automatically captures such effects and interactions through the model. Moreover, Fed-ComBat can be considered a better approach towards privacy preserving methods as it only shares model parameters, while d-ComBat and our included extension d-ComBat-GAM share full covariance and cross-covariance matrices (van Kesteren et al., 2019).

This work aims to contribute to the state-of-the-art by extending harmonization in multi-centric clinical studies to real decentralized setups that are compliant with data governance and sharing constraints. Additionally we contribute to the known challenge of data harmonization in the presence of non-linear covariate effects by removing the need to prior knowledge or assumptions to approximate such effects.

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), α_g acts as a global fixed intercept (i.e., the mean), while γ_{ig} indicates a random intercept that accounts for the site-specific shift. ε_{ijg} is a noise model that captures the variability of each phenotype $\varepsilon_{ijg} \sim \mathcal{N}(0, \sigma_g^2)$, and δ_{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). While the original model considered $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ to be a linear function of $\mathbf{x}$, and the nonlinear model considered $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ to be an univariate spline, this formulation treats $\phi(\mathbf{x}; \boldsymbol{\theta}_g)$ as a general function that can potentially be multivariable and nonlinear and parametrized by $\boldsymbol{\theta}_g$. This

generalization allows for the use of more complex approximating functions such as kernel-based models or neural networks as the multilayer perceptron (MLP), which is known to be an universal function approximator. However, it is essential to use cross-validation techniques to avoid overfitting and ensure the model generalizes well to unseen data.

Some constraints are defined during the fixed effect parameter estimation $(\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}, \hat{\sigma}_g)$:

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

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

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

A first constraint in Equation (3) is set to allow identifiability of the intercept parameters as explained by Johnson et al. (2007) (explained in practice in Appendix S3). 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 performed in three steps: i) maximum likelihood estimation (MLE) for parameters $\hat{\alpha}_g, \hat{\boldsymbol{\theta}}_g, \hat{\gamma}_{ig}$ (see Equation (2)) and of 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 $y_{ijg} \rightarrow z_{ijg} \sim \mathcal{N}(\gamma_{ig}, \delta^2_{ig})$ 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 γ_{ig} and δ^2_{ig} to iteratively estimate these parameters as in Equation (7) (Johnson et al., 2007, Sec. 3.2).

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

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

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

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

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

2.2. Federated ComBat: Fed-ComBat

Considering the formulation previously presented in Equation (1), the parameters $\alpha_g, \boldsymbol{\theta}_g$ and γ_{ig} can be optimized by minimizing an objective function F. As data is now siloed, it is only possible to have an evaluation of the cost function at each site F_i, thus defining the federated optimization problem as:

$$\arg \min_{\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}} F(\alpha_g, \boldsymbol{\theta}_g, \gamma_{ig}), \tag{9}$$

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

We rely on *Federated averaging* (FedAVG) to tackle this optimization problem (McMahan et al., 2017) by allowing each site to conduct partial optimization using stochastic gradient descent (SGD) locally on their data, followed by an aggregation step that combines the shared parameters of each site. The convergence of the resulting iterative process has been demonstrated for both IID and non-IID data distribution across clients (Li et al., 2019b).

The Fed-ComBat framework requires the following steps for harmonization: i) a federated standardization step to avoid scaling issues in gradient descent as proposed by Silva et al. (2019), ii) a federated estimation of $\alpha_g, \boldsymbol{\theta}_g$ and γ_{ig} as presented in Equation (10), and iii) a local estimation of the random effects following using EB as in Equations (6) and (7). A description of the steps followed in federated harmonization using Fed-ComBat is described in Algorithm 1 including standardization as preprocessing, local updates relying on SGD, global updates across sites of shared parameters using FedAVG and random effect estimation using Empirical Bayes.

2.3. Related works

2.3.1. d-ComBat

Chen et al. (2022) proposed a distributed version of ComBat that allows correction for batch effects in a distributed manner. This is achieved by taking advantage of the multivariate linear regression problem solved in ComBat, requiring to compute the product between the inverse of a covariance matrix and the data-targets cross-covariance. Both can be computed by aggregating the individual matrices provided by each site. This allows us to reformulate Equation (1) as done by (Chen et al., 2022):

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` γ_{ig}^* `and` δ_{ig}^* `using EB`

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

  `// Correct data`

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

$$
\underbrace{\begin{bmatrix} y_{11g} \\ \vdots \\ y_{ijg} \\ \vdots \\ y_{Sn_sg} \end{bmatrix}}_{\mathbf{y}_g} = \underbrace{\left[\begin{array}{ccc|ccc|c} x_{111} & \dots & x_{11C} & 1 & \cdots & 0 & 1 \\ \vdots & x_{ijc} & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{Sn_S1} & \dots & x_{Sn_sC} & 0 & \dots & 1 & 1 \end{array}\right]}_{\mathbf{X}} \underbrace{\left[\begin{array}{c} \hat{\theta}_{g1} \\ \vdots \\ \hat{\theta}_{gc} \\ \hline \hat{\gamma}_{1g} \\ \vdots \\ \hline \hat{\alpha}_g \end{array}\right]^{\intercal}}_{\hat{\mathbf{\Theta}}_g} + \boldsymbol{\delta}_g \boldsymbol{\varepsilon}_g \tag{11}
$$

Starting from the panel formulation where $\mathbf{y_g}$ is the stacked vector of all subjects across all sites i, and $\mathbf{X}$ is a design matrix containing the covariate effects to be preserved indexed by c in subject j from site i (x_{ijc}), an indicator matrix encoding the site, and a column of ones to capture α_g, we can estimate the augmented parameter matrix $\hat{\mathbf{\Theta}}_g$ using maximum likelihood. The estimator can be decomposed as a sum of the covariance and the cross-covariance matrices at each site, as shown in Equation 12. This formulation has also been used to distribute latent variable models, such as PCA and PLS, among multiple sites (Silva et al., 2019; Lorenzi et al., 2018). The authors have also suggested sharing a partial-eigen decomposition to mitigate the risk of data leakage, as shown in Equation 13 where the least squares problem is decomposed by $\mathbf{X}^{\intercal}\mathbf{X}$ which denotes the covariance matrix, $\mathbf{X}^{\intercal}\mathbf{y}_g$ the cross-covariance matrix for phenotype g, $\mathbf{Q}_i, \mathbf{\Lambda}_i$ the eigen decomposition for the covariance matrix in site i and, $\mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i$ the SVD decomposition for

the cross-covariance matrix in site i.

$$\hat{\mathbf{\Theta}}_g = \left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}_g = \left(\sum_{i=1}^{S}\mathbf{X}_i^\intercal\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{S}\mathbf{X}_i^\intercal\mathbf{y}_{ig}\right) \tag{12}$$

$$= \left(\sum_{i=1}^{S}\mathbf{Q}_i\mathbf{\Lambda}_i\mathbf{Q}_i^{-1}\right)\left(\sum_{i=1}^{S}\mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\intercal\right) \tag{13}$$

2.3.2. d-ComBat-GAM

Seeking a federated nonlinear approach, we here propose d-ComBat-GAM as an adaptation for d-ComBat for non-linear covariate effect preservation via generalized additive models (GAM) as proposed in its centralized version by Pomponio et al. (2020). This method, as d-ComBat, allows a fast estimation in a closed form by sharing either full or partial decomposition of the covariance and cross covariance matrices but transforming covariates in a design matrix using smooth functions such as polynomials or splines.

We can consider d-ComBat-GAM as a particular case of Fed-ComBat where $\phi(\cdot)$ becomes a linear combination parametrized by $\boldsymbol{\theta}$. Let $b(\cdot)$ be an arbitrary basis representation, the covariate function approximate is then defined as follows:

$$\phi(\mathbf{x}_{ij};\boldsymbol{\theta}_g) = \sum_{c=1}^{C}\sum_{k=1}^{K_c}\theta_{gk}b_{gk}(x_{ijc}) \quad \text{where} \quad \boldsymbol{\theta}_g = [\theta_{g1},\ldots,\theta_{gK_c}] \tag{14}$$

Note that Equation (14) expresses a linear combination of parameters, hence the same procedure as in Section 2.3.1 can be followed to estimate the fixed effects. Although polynomial decomposition is straightforward as it is the equivalent of transforming and appending the variable to a design matrix, splines require a set of defined control points based on the covariate range. We rely on federated standardization to map the nonlinear covariate

to a normalized distribution where the range is known and thus the control points of the splines can be predefined.

3. Materials

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

3.1. Synthetic data

We created simulated datasets under three main scenarios: i) when cohorts have similar distributions for covariates and sample size effects (IID), ii) when covariate distributions slightly overlap but follow different patterns, and when there are differences in sample sizes (mid-non-IID), and iii) when there is no overlap between covariate distributions between cohorts and there are evident differences in sample sizes (non-IID). An example of IID and non-IID covariate distributions on synthetic data is shown in Figure 1. We simulated ComBat parameters following the methodology outlined in Reynolds et al. (2022), using the graphical model depicted in Figure 2. The primary difference between our approach and that of Reynolds et al. (2022) was the model used to relate the covariates and phenotypes ϕ. We used a nonlinear function in this step, which was employed as the unbiased target function to be learned after the harmonization process for the synthetically biased phenotypes.

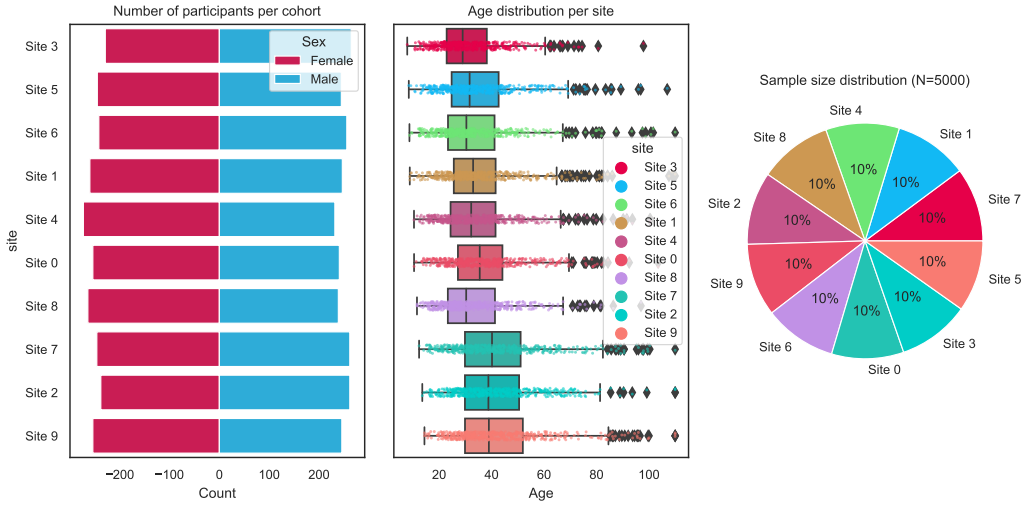
Sex was generated by sampling from a Bernoulli($p_i^{(\text{sex})}$) distribution, which modulated sex proportion distribution across sites. Age was drawn from a $\mathcal{N}(\mu_i^{(\text{age})}, {\sigma_i^{(\text{age})}}^2)$ distribution, where $\mu_i^{(\text{age})}$ and $\sigma_i^{(\text{age})}$ were sampled from a Uniform(a, b) distribution, with heterogeneity modulated by the width of the distribution $(b - a)$. Sample size heterogeneity was simulated using a concentration model $p_i \sim \text{Dir}(\boldsymbol{\alpha})$, as previously suggested for measuring the effects of heterogeneity in Federated Learning by Hsu et al. (2019).

As shown in Figure 3 and explained by multiple authors in the review by Zhu et al. (2021), non-IID scenarios consistently pose the greatest challenges for federated approaches. Therefore, we chose to focus on non-IID setups to provide a more realistic and comprehensive evaluation.

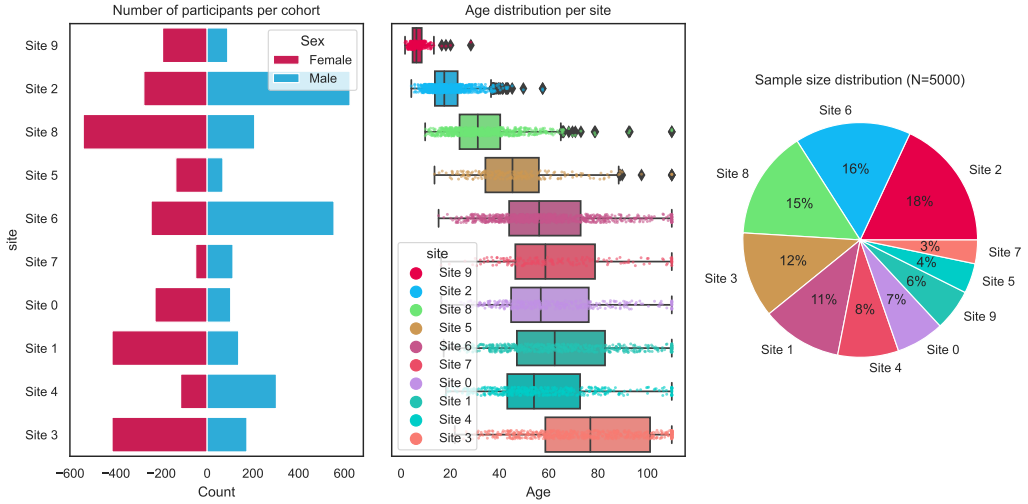
3.2. Brain MRI-Data

We evaluated our method on a cross-sectional cohort of 7265 participants from nine public studies, utilizing MRI-derived phenotypes from baseline structural T1-weighted magnetic resonance imaging (MRI).

Participants were categorized by diagnosis as follows: 3992 cognitively normal (CN), of which 731 of them were considered at risk of cognitive impairment from the A4 study (Sperling et al., 2014); 1065 were identified as having mild cognitive impairment (MCI) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Weiner et al., 2013) and the Australian Imaging, Biomarker and Lifestyle Flagship Study of Aging (AIBL) (Ellis et al., 2009); 998 had been diagnosed with Autism Spectrum Disorder (ASD) from ABIDE-I and ABIDE II (Di Martino et al., 2014); and 538 were patients diagnosed with Alzheimer's disease (AD) from ADNI and Open Access Series of Imaging Studies (OASIS) dataset (LaMontagne et al., 2019) and, 686



(a)



(b)

Figure 1: Population distribution of synthetic data across cohorts in two different scenarios. The left panel shows the sex distribution, the middle panel shows the age distribution, and the right panel shows the sample size distribution. Panel 1a shows a homogeneous and IID population, while panel 1b shows a non-IID distribution of covariates, which is more commonly observed and used in this work.

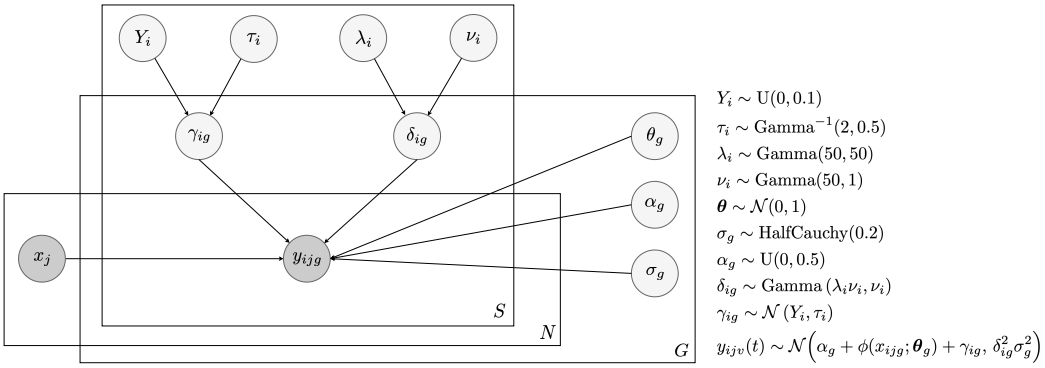


Figure 2: Graphical model used to generate synthetic data. The shaded circles indicate observed measurements, including covariates and imaging feature values, while unshaded circles represent latent parameters.

participants diagnosed in multiple phases of Parkinson's disease from the Parkinson's Disease Biomarkers Program (PDPB) (Rosenthal et al., 2016) and the Parkinson's Progression Markers Initiative (PPMI) (Marek et al., 2018). More detailed demographic information can be found in Table 1. Covariate distribution across the multiple cohorts used in this work can be better observed in Figure 3, showing that when using multiple cohorts, it's very likely to have non-IID distributions.

3.2.1. Image processing

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.

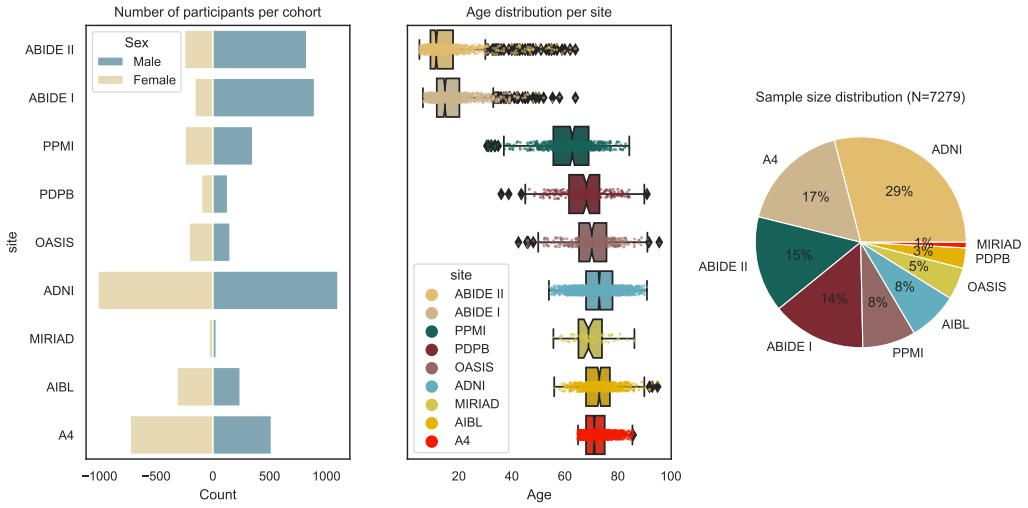


Figure 3: Population pyramid representing the demographics of the real data used in this study. The left panel shows the distribution of sex, the middle panel shows the distribution of age, and the right panel shows the sample size distribution. Cohorts were sorted by ascending median age. The demographics for the population are described in Table 1. The figure shows a clear non-IID distribution, similar to the distribution observed in Figure 1b.

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

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	**CN**	728	515	71.3 ± 4.5 [65.0 - 85.7]	72.9 ± 5.1 [65.0 - 85.7]
ABIDE I	**Autism**	59	434	16.3 ± 8.1 [8.1 - 45.0]	17.4 ± 8.6 [7.0 - 64.0]
	CN	98	446	15.4 ± 6.6 [7.8 - 46.0]	17.4 ± 7.6 [6.5 - 48.0]
ABIDE II	**Autism**	74	416	13.1 ± 8.1 [5.2 - 54.0]	15.0 ± 9.3 [5.1 - 62.0]
	CN	175	397	13.5 ± 7.4 [5.9 - 46.6]	15.6 ± 10.0 [5.9 - 64.0]
ADNI	**AD**	156	190	73.6 ± 7.8 [55.0 - 91.0]	75.0 ± 7.6 [55.0 - 90.0]
	CN	444	345	71.8 ± 6.2 [55.0 - 90.0]	74.0 ± 6.1 [60.0 - 90.0]
	MCI	406	555	71.5 ± 7.8 [55.0 - 88.0]	73.6 ± 7.2 [54.0 - 90.0]
AIBL	**AD**	34	25	75.6 ± 8.1 [58.0 - 93.0]	72.7 ± 6.9 [60.0 - 83.0]
	CN	237	169	71.9 ± 6.0 [60.0 - 89.0]	73.0 ± 6.3 [60.0 - 90.0]
	MCI	44	46	75.8 ± 8.4 [56.0 - 95.0]	74.1 ± 5.6 [64.0 - 87.0]
MIRIAD	**AD**	22	17	69.5 ± 6.3 [58.1 - 80.2]	69.4 ± 7.5 [55.7 - 86.1]
	CN	11	12	66.3 ± 4.9 [58.8 - 73.8]	73.6 ± 7.2 [63.5 - 86.3]
OASIS	**AD**	45	41	75.6 ± 6.8 [62.9 - 95.6]	75.2 ± 7.3 [60.5 - 91.7]
	CN	162	109	67.6 ± 8.6 [45.7 - 88.8]	69.1 ± 8.3 [42.5 - 86.2]
PDPB	**PD**	97	129	67.8 ± 9.1 [50.7 - 90.0]	67.2 ± 9.7 [36.0 - 91.0]
PPMI	**CN**	51	84	59.1 ± 11.6 [31.0 - 81.9]	60.8 ± 11.7 [30.6 - 82.8]
	GenCohort PD	29	27	65.2 ± 8.0 [51.7 - 81.2]	63.9 ± 9.3 [32.2 - 78.6]
	GenCohort Unaff	52	35	61.1 ± 8.2 [33.7 - 84.3]	61.3 ± 7.6 [46.8 - 75.0]
	PD	94	164	61.2 ± 10.2 [33.5 - 81.7]	62.2 ± 9.4 [34.8 - 82.9]
	Prodromal	2	14	70.1 ± 4.3 [67.1 - 73.2]	69.0 ± 6.7 [61.7 - 82.5]
	SWEDD	15	26	61.7 ± 8.9 [46.8 - 77.6]	61.9 ± 9.9 [39.2 - 77.0]

4. Results

Two versions of Fed-ComBat were used for comparison: a first one defining a linear model ($\phi(\mathbf{x};\boldsymbol{\theta}) = \mathbf{x}^\intercal\boldsymbol{\theta}$) to show equivalence with respect to ComBat and d-ComBat, and a second one defining ϕ as a multi-layer perceptron model (MLP) for nonlinear effect preservation. These two versions of Fed-ComBat were tested jointly with our extended proposal d-ComBat-GAM against their two centralized versions: ComBat[1] (a.k.a. NeuroComBat), ComBat-GAM[2] and the only reported distributed version d-ComBat-GAM[3].

4.1. Synthetic data

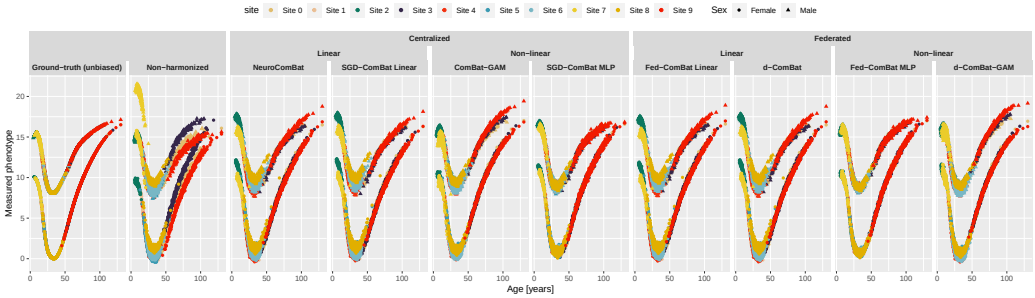
Reconstruction of an unbiased phenotype function depending on age and sex after harmonizing using different methods is shown in Figure 4a. As observed, centralized ComBat, d-ComBat, and Fed-ComBat in its linear version follow similar trends. It can also be observed that for nonlinear covariate effects linear correction can even induce biases due to its limitation to capture complex relationships between covariates and phenotypes (see Non-harmonized vs. NeuroCombat in Figure 4a).

While both ComBat-GAM and our distributed adaptation d-ComBat-GAM demonstrate improved correction for nonlinear covariate effects compared to ComBat (as shown in Figure 4a), it is worth noting that d-ComBat-GAM requires a priori definition of covariate interactions. In cases where these interactions are not defined, the performance of d-ComBat-GAM is

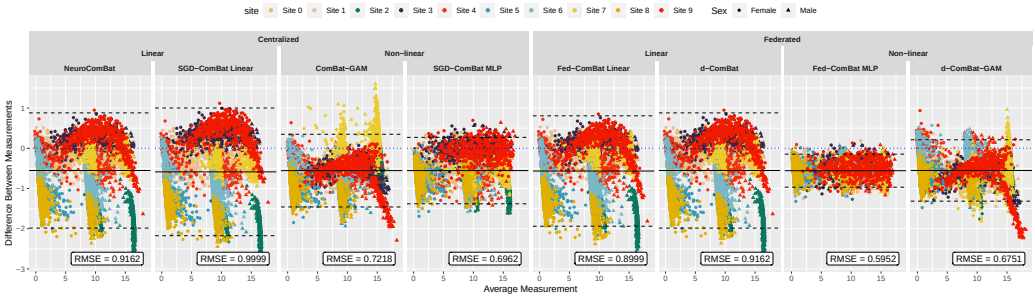
[1] `https://github.com/Jfortin1/neuroCombat`
[2] `https://github.com/rpomponio/neuroHarmonize`
[3] `https://github.com/andy1764/Distributed-ComBat`

reduced to that of ComBat. A better reconstruction of the target unbiased trajectories is achieved with Fed-ComBat using a two-layer perceptron neural network with 100 hidden units and with no hyperparameter tuning. Hyperparameter tuning was out of the scope of this work. Root mean square errors (RMSE) evaluating the reconstruction quality is presented in Table 2 for all centralized and federated methods.

Bland-Altman plots contrasting the difference between the different harmonization methods and the ground truth are presented in Figure 4b highlighting that there is less error variability (σ_ε) in estimating the unbiased function after harmonization for nonlinear methods and, particularly for Fed-ComBat in terms of the Root mean square error (RMSE = 0.5952) compared to the second federated best d-ComBat-GAM (RMSE = 0.6751).



(a) Reconstruction of an unbiased groundtruth function.



(b) Bland-Altmann plots and root mean square errors (RMSE) contrasting the different centralized and federated harmonization methods against the groundtruth.

Figure 4: Qualitative comparison across harmonization methods evaluating the quality of harmonization for a simulated phenotype with a nonlinear relationship with age and following a different trajectory per group (males and females). Results are shown for the extreme non-IID covariate distribution is depicted in Figure 1b. This can be seen as certain centers contain extremely young or old cohorts (tails). (a) Expected result after harmonization: groundtruth (unbiased), worst case scenario: biased data (non-harmonized), and harmonized phenotypes using the different ComBat methods considered and proposed in this work. As shown, ComBat could even insert biases when data is too heterogeneous. Fed-ComBat preserves better the real trajectories than the linear approaches and ComBat-GAM without the need of defining where the nonlinearities may be. Fed-ComBat using a multi-layer perceptron (MLP) shows the best reconstruction with an improvement of 35% in RMSE with respect to d-ComBat and 12% with respect to d-ComBat-GAM.

Setup	Approach	Method	$\lvert\bar{\varepsilon}\rvert$ (MAE)	σ_ε	RMSE
Federated	Non-linear	Fed-ComBat MLP	0.55710	**0.20960**	**0.59520**
Federated	Non-linear	d-ComBat-GAM	**0.55160**	0.38930	0.67510
Centralized	Non-linear	SGD-ComBat MLP	0.55420	0.42150	0.69620
Centralized	Non-linear	ComBat-GAM	0.55530	0.46120	0.72180
Federated	Linear	Fed-ComBat Linear	0.56540	0.70020	0.89990
Centralized	Linear	NeuroComBat	0.55230	0.73110	0.91620
Federated	Linear	d-ComBat	0.55230	0.73110	0.91620
Centralized	Linear	SGD-ComBat Linear	0.58640	0.81000	0.99990

Table 2: Quantitative results for Figure 4a in terms of mean absolute error between methods and the target hidden unbiased trajectories ($|\bar{\varepsilon}|$), its standard deviation (σ_ε) and root mean square error (RMSE). Highlighted values correspond to the best metric across different methods.

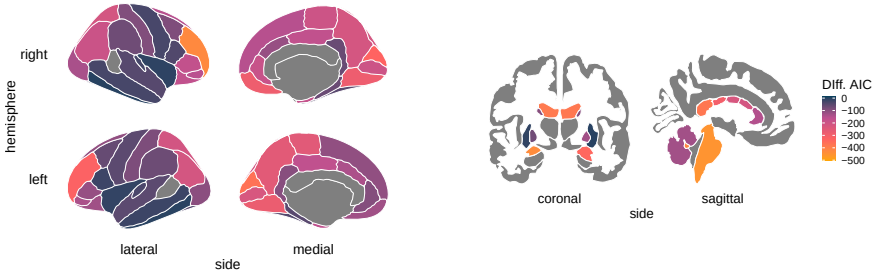
4.2. Brain MRI-data

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 firstly centralized and harmonized accounting for study as the batch source using ComBat. The criterion used to evaluate the presence of nonlinearities was the difference in the Akaike Information Criterion (AIC) between the GAM and the linear model. When the difference is negative, it indicates that the GAM is a better fit, and the magnitude of the difference indicates by how much. Figure 5b 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. The data was controlled for sex, diagnosis, ICV, and diagnosis group, and age was considered an independent variable.

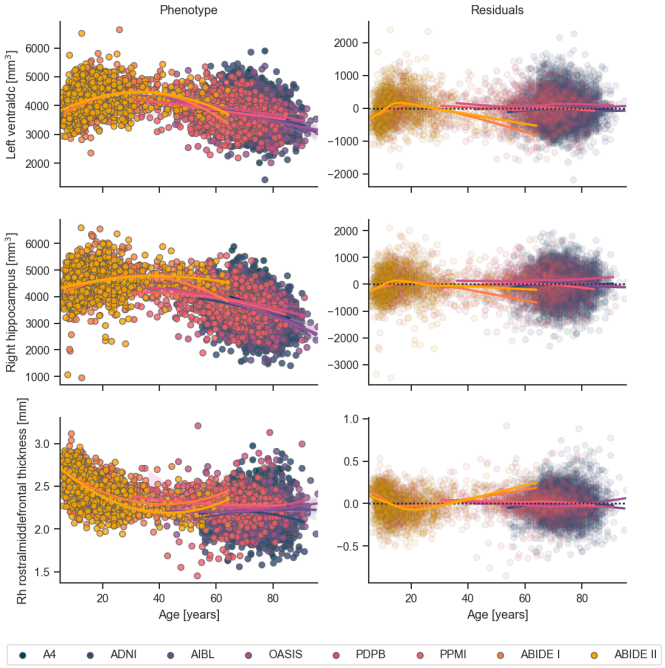
Figure 5a shows the top three regions that are better explained by a nonlinear model of age in terms of AIC, as well as the residuals after the regression illustrating remaining effects or trends. More detailed results on the AICs differences is presented in Table S3.

Then data was decentralized and we illustrate the ages trajectories by regrouping participants in two trajectories: cognitively normal were labeled as "Healthy controls" while those diagnosed with a particular neurological disorder were labeled as "Atypical". The harmonized phenotypes were obtained using as linear centralized method: ComBat, nonlinear centralized: d-ComBat, linear federated: d-ComBat and our adaptation Fed-ComBat using a linear model, and nonlinear federated: our extension d-ComBat-GAM and Fed-ComBat using an MLP with two layers and 100 hidden units.

A GAM was used to estimate these trajectories after each method controlling for Sex and ICV. Resulting trajectories are illustrated in Figure 6. The results show that our proposals Fed-ComBat Linear and d-ComBat-GAM produce consistent results with respect to ComBat, d-ComBat, and ComBat-GAM. After the integration of the MLP as covariate effect approximating function, both centralized and federated show similar trajectory derivation. The structure of the trajectories is preserved with the main difference being the exacerbation or attenuation of certain effects like the higher rate of deterioration in thickness for the right rostral middle frontal cortex in the young population presented in Figure 6 bottom right panel using Fed-ComBat MLP, and a less attenuated effect in the elder cohort on the right hippocampal volume confirming not only the viability of global harmonization without sharing data, but also doing it with a more privacy-aware



(a) 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.



(b) Results of regression using a GAM on the top 3 regions with the most discrepant AIC values on the left and residual plots illustrating goodness of fit on the right. Top: Ventral diencephalon, middle: Right hippocampus, bottom: Right hemisphere rostral middle frontal gyrus.

Figure 5: Evidence of nonlinear effects present in used cohorts. Subjects were harmonized by cohort using ComBat (Fortin et al., 2018).

approach using FL and integrating more complex models such as networks.

Trajectories traced after harmonizing with ComBat-GAM and d-ComBat-GAM suggest an increase in right hippocampal volume which does not seem to be supported by the literature and seems more like a residual site effect.

Another important point to remark is the inflection points in age where right hippocampal volume reduction begins. While the non-harmonized data shows that reduction starts to occur in the late 50s, linear approaches exacerbate this effect in these cohorts, suggesting a decline after 25 years old in cognitively normal populations, which can be questionable with respect to what has been reported so far in this region (Nobis et al., 2019; De Francesco et al., 2021; Liu et al., 2021).

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). This results are almost identical and do not show residual batch effects after harmonization. This is due to the nature of the formulation of each ComBat model, suggesting this visualization not to be the best for qualitative evaluation.

Further analysis and evaluation on datasets and cohorts specially where data sharing is prohibited is required. As pointed by Gebre et al. (2023) further exploration is needed in harmonization methods as some of the central-

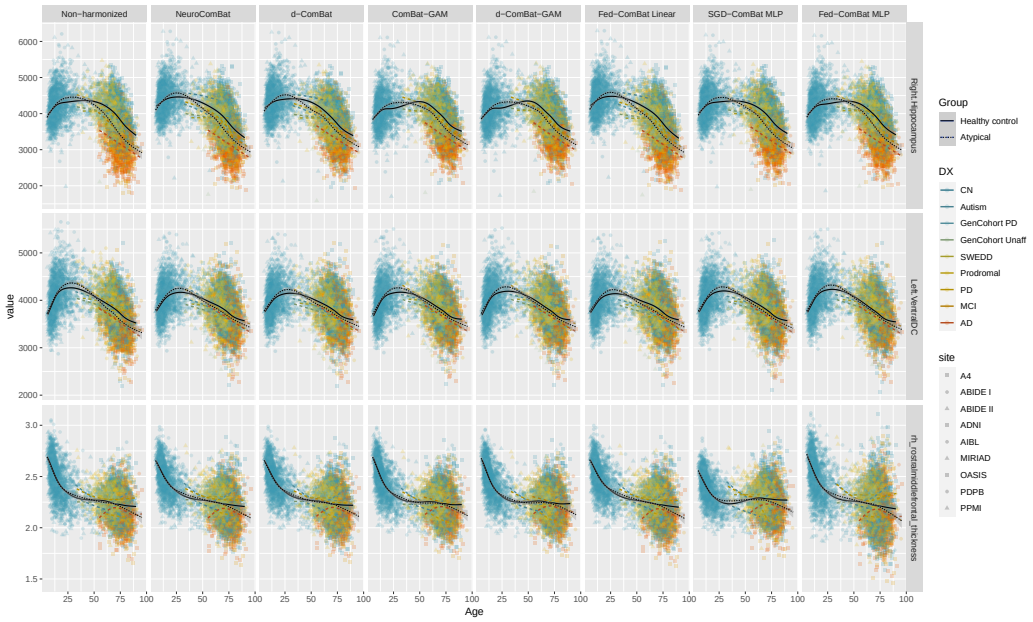


Figure 6: Comparison of age trajectories for three brain regions: right hippocampus (top row), left ventral diencephalon (middle row), and right rostral middle frontal cortex thickness (bottom row). The columns represent conditions under which the data is shown: non-harmonized (first column), and harmonized using NeuroComBat (Fortin et al., 2017), d-ComBat (Chen et al., 2022), ComBat-GAM (Pomponio et al., 2020), d-ComBat-GAM (us), Fed-ComBat Linear† (federated equiv. to d-ComBat but SGD based), SGD-ComBat MLP† (centralized using MLP) and Fed-ComBat MLP† (decentralized). Trajectories were adjusted for sex and ICV. Fed-ComBat MLP shows to be similar to ComBat maintaining trajectory structure while offering a more privacy-aware approach given that data is kept decentralized and no full covariance matrices are being shared in this approach and the harmonization is sill being performed globally. †: Proposed in this work.

ized methods evaluated in this study may result in non-satisfactory results.

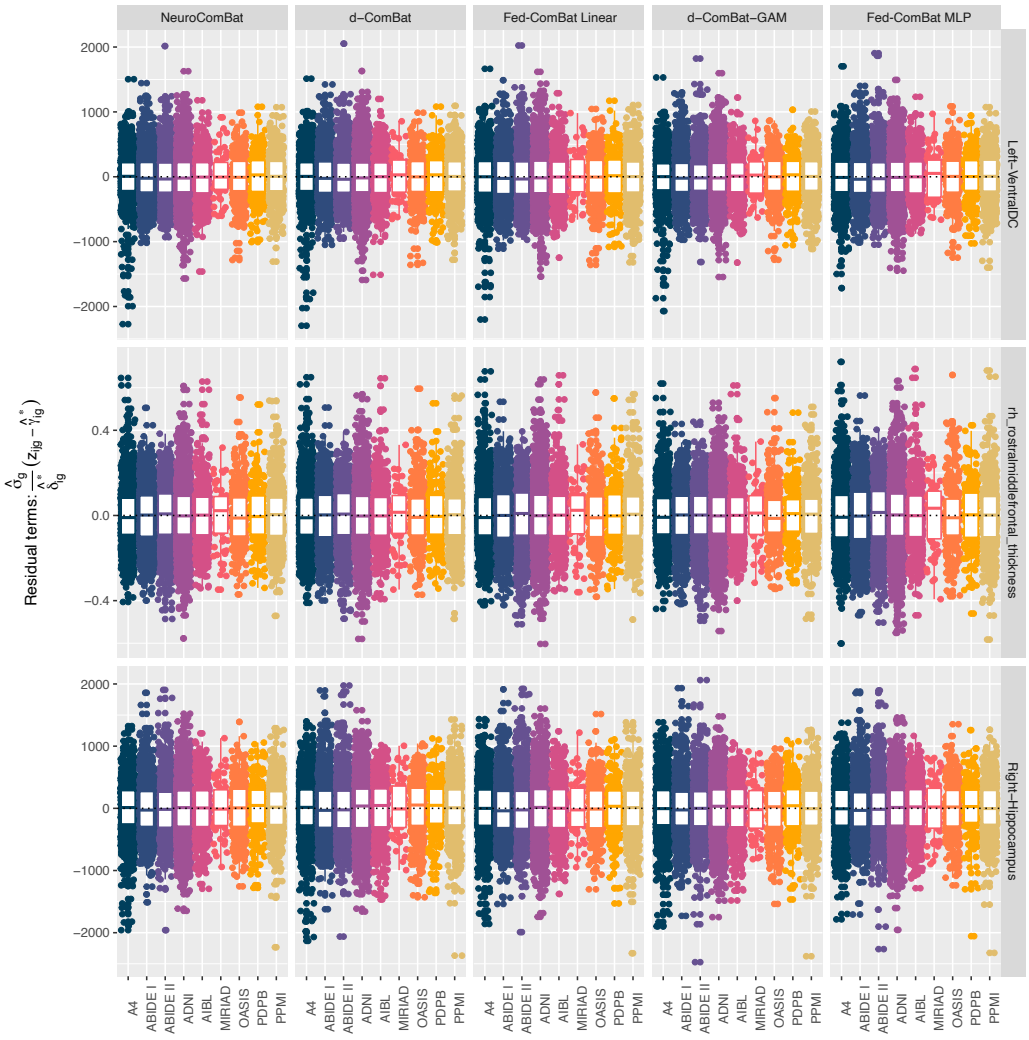


Figure 7: Residual terms after harmonization for top three regions discusses in Figure 5a for federated methods. Additional comparisons using all the methods are presented in Appendix S2.

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 ComBat and d-ComBat. 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 multilayer perceptron (MLP) versions, is comparable to that of centralized methods. 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 across such, 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.

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.

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. 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., 2020a) 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 worth-

while 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.

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

8. 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 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 Alzheimer's Disease Neuroimaging Initiative

(ADNI) database (https://adni.loni.usc.edu/). Data used in this article were provided by 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. Data used in this article were obtained from the Autism Brain Imaging Data Exchange (ABIDE) II database (http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html). 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 MIRIAD database. 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/). Data used in the preparation of this article were obtained from the Pre-symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease (A4) study.

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.

Bukaty, P., 2019. The California Consumer Privacy Act (CCPA): An implementation guide. IT Governance Publishing. URL: `http://www.jstor.org/stable/j.ctvjghvnn`.

Calzada, I., 2022. Citizens' data privacy in china: The state of the art of the personal information protection law (pipl). Smart Cities 5, 1129–1150.

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.

Cheung, A.T., Nasir-Moin, M., Kwon, Y.J.F., Guan, J., Liu, C., Jiang, L., Raimondo, C., Chotai, S., Chambless, L., Ahmad, H.S., et al., 2023. Methods and impact for using federated learning to collaborate on clinical research. Neurosurgery 92, 431–438.

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

De Francesco, S., Galluzzi, S., Vanacore, N., Festari, C., Rossini, P.M., Cappa, S.F., Frisoni, G.B., Redolfi, A., 2021. Norms for automatic estimation of hippocampal atrophy and a step forward for applicability to the italian population. Frontiers in Neuroscience , 786.

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.

Di Martino, A., O'connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., et al., 2017. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. Scientific data 4, 1–15.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19, 659–667.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., et al., 2009. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. International psychogeriatrics 21, 672–687.

European Commission, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). URL: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

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. Harmonization 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.

Hsu, T.M.H., Qi, H., Brown, M., 2019. Measuring the effects of non-identical data distribution for federated visual classification. arXiv preprint arXiv:1909.06335 .

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

van Kesteren, E.J., Sun, C., Oberski, D.L., Dumontier, M., Ippel, L., 2019. Privacy-preserving generalized linear models using distributed block coordinate descent. arXiv preprint arXiv:1911.03183 .

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 .

LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al., 2019. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv , 2019–12.

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

Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., et al., 2019a. Privacy-preserving federated brain tumour segmentation, in: International workshop on machine learning in medical imaging, Springer. pp. 133–141.

Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020b. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. Medical Image Analysis 65, 101765.

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

Liu, Y., Meng, J., Wang, K., Zhuang, K., Chen, Q., Yang, W., Qiu, J., Wei, D., 2021. Morphometry of the hippocampus across the adult life-span in patients with depressive disorders: Association with neuroticism. Brain Topography 34, 587–597.

Lorenzi, M., Altmann, A., Gutman, B., Wray, S., Arber, C., Hibar, D.P., Jahanshad, N., Schott, J.M., Alexander, D.C., Thompson, P.M., et al., 2018. Susceptibility of brain atrophy to trib3 in alzheimer's disease, evidence from functional prioritization in imaging genetics. Proceedings of the National Academy of Sciences 115, 3162–3167.

Mahlool, D.H., Abed, M.H., 2022. Distributed brain tumor diagnosis using a federated learning environment. Bulletin of Electrical Engineering and Informatics 11, 3313 – 3321. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139133360&doi=10.11591%2feei.v11i6.4131&partnerID=40&md5=1ad3123e9db42a98e13e51bd26fdcbc1, doi:10.11591/eei.v11i6.4131. cited by: 0; All Open Access, Gold Open Access, Green Open Access.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M., 2013. Miriad—public release of a multiple time point alzheimer's mr imaging dataset. NeuroImage 70, 33–36.

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C.S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C.M., Trojanowski, J.Q., et al., 2018. The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort. Annals of clinical and translational neurology 5, 1460–1477.

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.

Nobis, L., Manohar, S.G., Smith, S.M., Alfaro-Almagro, F., Jenkinson, M., Mackay, C.E., Husain, M., 2019. Hippocampal volume across age: Nomograms derived from over 19,700 people in uk biobank. NeuroImage: Clinical 23, 101904. URL: https://www.sciencedirect.com/science/article/pii/S2213158219302542, doi:https://doi.org/10.1016/j.nicl.2019.101904.

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 .

Rosenthal, L.S., Drake, D., Alcalay, R.N., Babcock, D., Bowman, F.D., Chen-Plotkin, A., Dawson, T.M., Dewey Jr, R.B., German, D.C., Huang,

X., et al., 2016. The ninds parkinson's disease biomarkers program. Movement Disorders 31, 915–923.

Schwarz, C.G., Kremers, W.K., Wiste, H.J., Gunter, J.L., Vemuri, P., Spychalla, A.J., Kantarci, K., Schultz, A.P., Sperling, R.A., Knopman, D.S., et al., 2021. Changing the face of neuroimaging research: Comparing a new mri de-facing technique with popular alternatives. NeuroImage 231, 117845.

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.

Silva, S., Gutman, B.A., Romero, E., Thompson, P.M., Altmann, A., Lorenzi, M., 2019. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data, in: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), IEEE. pp. 270–274.

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.

Sperling, R.A., Rentz, D.M., Johnson, K.A., Karlawish, J., Donohue, M.,

Salmon, D.P., Aisen, P., 2014. The a4 study: stopping ad before symptoms begin? Science translational medicine 6, 228fs13–228fs13.

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.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al., 2013. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimer's & Dementia 9, e111–e194.

Zhu, H., Xu, J., Liu, S., Jin, Y., 2021. Federated learning on non-iid data: A survey. Neurocomputing 465, 371–390.

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

Santiago Silva[a], Neil P. Oxtoby[c], Andre Altmann[b], Marco Lorenzi[a,], for the Alzheimer's Disease Neuroimaging Initiative

[a] *Université Côte d'Azur and INRIA, EPIONE team, Sophia Antipolis, France*
[b] *Centre for Medical Image Computing (CMIC), Department of Medical Physics and Biomedical Engineering, University College London, London, UK*
[c] *UCL Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK*

Contents

S1	**Comparing centralized and Fed-ComBat's formulation**	**s1**
S2	**Harmonization on real data**	**s3**
S2.1	Evidence of non linear effects on brain phenotypes	s3
S2.2	Residual ComBat effects on real data.	s3
S3	**Identifiability of ComBat parameters**	**s6**
S4	**CRediT author statement**	**s7**

S1. Comparing centralized and Fed-ComBat's formulation

As the random effects (batch effects) in this work are assumed to be site-specific, the model's design matrix $\mathbf{X}$ is constructed under a linear assump-

Email address: marco.lorenzi@inria.fr (Marco Lorenzi)

tion, jointly optimizing the parameters $\hat{\alpha}_g$ and $\hat{\gamma}_{ig}$. However, for nonlinear scenarios, the model needs to be reformulated, as previously done by Pomponio et al. (2020). To provide a more comprehensive understanding of the optimization problem, we propose the following generalized approach:

$$y_{ijg} = \phi(\mathbf{x}_{ij}; \boldsymbol{\theta}_g) + \mathbf{z}_{ij}\boldsymbol{\beta}_g + \delta_{ig}\varepsilon_{ijg} \tag{A.1}$$

Where with respect to the model in Equation (1), $\mathbf{Z} := \{\mathbf{z}_{ij}\}_{\forall_{i,j}}$ is an indicator matrix as indicated in Equation (A.2) on the batch effects that allows the elements of $\boldsymbol{\beta}_g$ be composites of α_g and γ_{ig} as shown in Appendix S1.

$$\mathbf{z}_{ij_{(1,S)}} := \begin{cases} 1 & \text{if } \mathbf{x}_{ij} \in \text{batch i,} \\ 0 & \text{if } \mathbf{x}_{ij} \notin \text{batch i.} \end{cases} \tag{A.2}$$

$$\boldsymbol{\beta}_g = \alpha_g + \begin{pmatrix} \gamma_{1g} \\ \vdots \\ \gamma_{Sg} \end{pmatrix} \tag{A.3}$$

$$\text{s.t.} \quad \mathbb{E}[\gamma_i] = \textstyle\sum_{i=1}^{S} \frac{n_i}{N}\gamma_{ig} = 0 \quad \forall_{g\in\{1,\ldots,G\}} \tag{A.4}$$

Which with the constraint in Eqn A.4 the estimation of α_g and γ_{ig} from β_g becomes straightforward:

$$\alpha_g = \mathbb{E}[\boldsymbol{\beta}_g] = \sum_{i=0}^{S} p_i\beta_{ig} = \cancelto{1}{\sum_{i=0}^{S} \frac{n_i}{N}}\alpha_g + \cancelto{0}{\mathbb{E}[\gamma_{ig}]} \tag{A.5}$$

$$\gamma_{ig} = \beta_{ig} - \alpha_g \tag{A.6}$$

S2. Harmonization on real data

S2.1. Evidence of non linear effects on brain phenotypes

Table S3 shows the difference in Akaike information criterion (AIC) between a GAM and a GLM approximating the phenotypes (brain measures) as proposed by Pomponio et al. (2020) (B-Splines for approximating age with DOF=10 and degree 3) as a function of Age, Sex, ICV and group (diagnosis). Negative values indicate a better fit using GAMs and hence the presence of nonlinear age effect.

S2.2. Residual ComBat effects on real data.

Figure S1 presents the residual plots obtained after fitting a comprehensive model to the harmonized MRI-derived phenotype data, incorporating variables such as age, sex, and group. These residual plots provide insightful visual representations of the model's performance. Notably, the observed absence of pronounced shifts or variance scaling in all the models suggests that a linear correction approach proves to be more than adequate for addressing the particular use case (see Section 5).

Region	ΔAIC	Region	ΔAIC	Region	ΔAIC
rh_rostralmiddlefrontal_thickness	-737.827	**rh_lateralorbitofrontal_thickness**	-294.661	**rh_fusiform_thickness**	-128.352
Right-Hippocampus	-666.327	**rh_superiorfrontal_thickness**	-284.615	**lh_caudalmiddlefrontal_thickness**	-126.349
Left-VentralDC	-631.412	**rh_lateraloccipital_thickness**	-284.338	**CC_Anterior**	-124.647
Left-Hippocampus	-588.296	**rh_posteriorcingulate_thickness**	-277.702	**lh_fusiform_thickness**	-124.049
lh_precuneus_thickness	-586.472	**rh_pericalcarine_thickness**	-277.491	**rh_isthmuscingulate_thickness**	-111.719
Right-VentralDC	-577.552	**CC_Mid_Posterior**	-276.176	**rh_parsopercularis_thickness**	-106.362
lh_rostralmiddlefrontal_thickness	-570.793	**lh_lateralorbitofrontal_thickness**	-258.546	**rh_parahippocampal_thickness**	-105.893
lh_cuneus_thickness	-544.505	**rh_parsorbitalis_thickness**	-246.652	**Right-Cerebellum-Cortex**	-100.931
Left-Lateral-Ventricle	-530.619	**rh_rostralanteriorcingulate_thickness**	-243.986	**Right-Pallidum**	-94.757
rh_cuneus_thickness	-515.156	**lh_parsorbitalis_thickness**	-230.732	**rh_middletemporal_thickness**	-87.857
Right-Lateral-Ventricle	-514.104	**Left-Caudate**	-229.629	**lh_parahippocampal_thickness**	-83.225
rh_precuneus_thickness	-505.154	**Right-Caudate**	-226.743	**lh_parsopercularis_thickness**	-67.404
rh_lingual_thickness	-489.542	**CC_Central**	-217.088	**lh_transversetemporal_thickness**	-67.285
Brain-Stem	-451.604	**lh_parstriangularis_thickness**	-215.885	**lh_superiortemporal_thickness**	-63.792
lh_paracentral_thickness	-443.599	**lh_lateraloccipital_thickness**	-213.35	**lh_middletemporal_thickness**	-59.581
lh_lingual_thickness	-443.392	**CC_Mid_Anterior**	-209.57	**rh_superiortemporal_thickness**	-57.407
lh_inferiorparietal_thickness	-439.908	**lh_superiorfrontal_thickness**	-206.472	**rh_transversetemporal_thickness**	-56.076
lh_superiorparietal_thickness	-424.113	**rh_caudalmiddlefrontal_thickness**	-206.01	**lh_insula_thickness**	-54.416
rh_superiorparietal_thickness	-420.171	**rh_entorhinal_thickness**	-194.472	**rh_insula_thickness**	-53.683
rh_medialorbitofrontal_thickness	-417.312	**lh_caudalanteriorcingulate_thickness**	-191.647	**lh_precentral_thickness**	-51.688
Left-Amygdala	-409.543	**lh_medialorbitofrontal_thickness**	-189.768	**lh_inferiortemporal_thickness**	-48.254
Right-Cerebellum-White-Matter	-392.708	**lh_entorhinal_thickness**	-176.719	**Left-Pallidum**	-40.941
rh_inferiorparietal_thickness	-381.635	**lh_isthmuscingulate_thickness**	-168.909	**rh_precentral_thickness**	-37.279
rh_paracentral_thickness	-358.113	**rh_supramarginal_thickness**	-161.94	**rh_inferiortemporal_thickness**	-37.05
Right-Amygdala	-324.232	**lh_postcentral_thickness**	-157.198	**Left-Putamen**	-34.233
CC_Posterior	-323.508	**rh_postcentral_thickness**	-154.755	**Right-Putamen**	-24.207
rh_parstriangularis_thickness	-310.166	**lh_supramarginal_thickness**	-148.167		
lh_posteriorcingulate_thickness	-306.379	**rh_caudalanteriorcingulate_thickness**	-144.195		
lh_pericalcarine_thickness	-305.481	**lh_rostralanteriorcingulate_thickness**	-128.495		

Table S3: Difference of Akaike information criterion (AIC) between a nonlinear and a linear model using age as the exogenous variable across cortical thickness measures. $\Delta\text{AIC} < 0$ suggest that the nonlinear model is a best fit hence suggesting non linear effects of age in such negative differences.

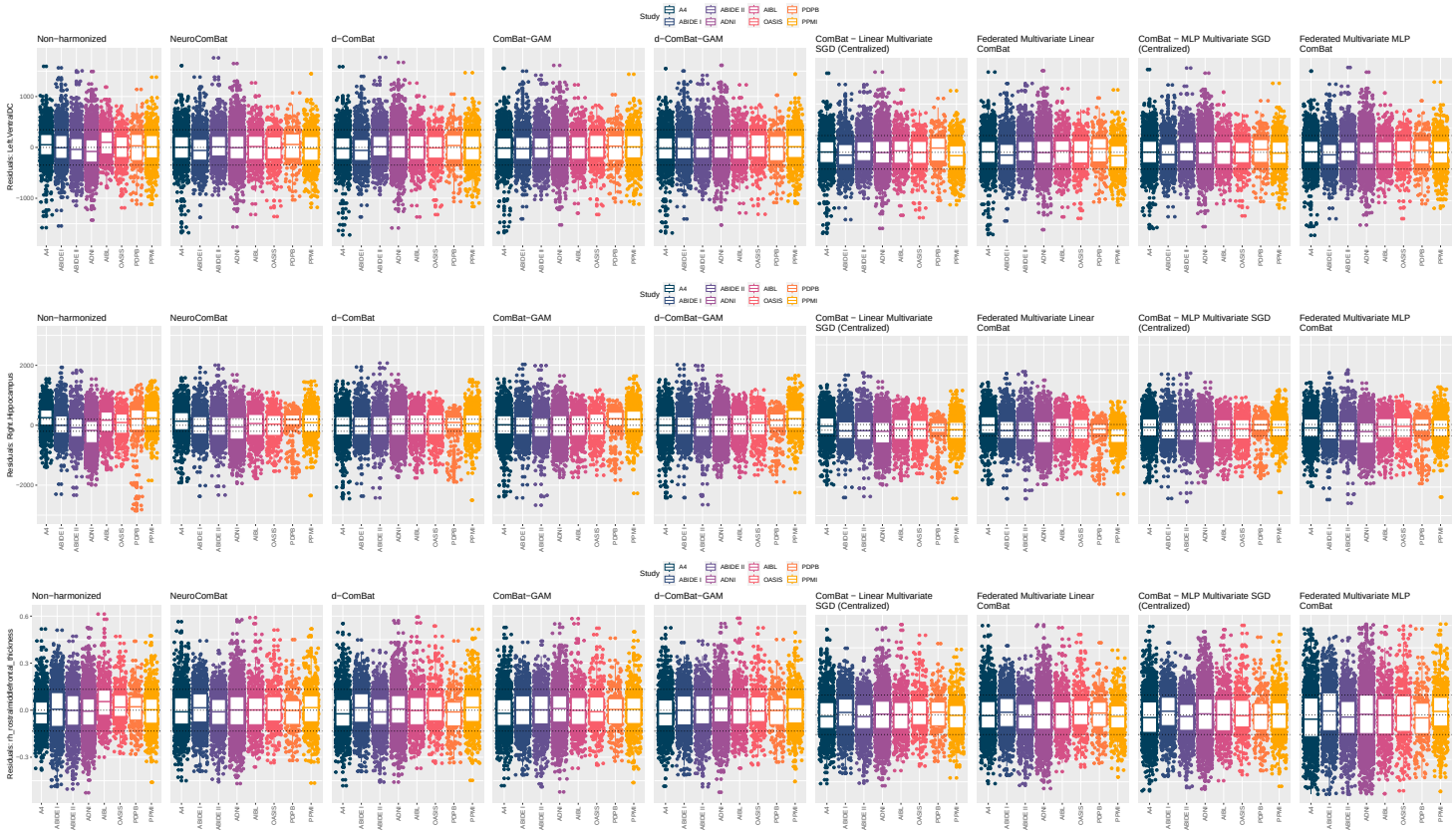


Figure S1: Figure 7 as proposed by Pomponio et al. (2020) for all the methods evaluated in this work and top regions in Figure 5a.

S3. Identifiability of ComBat parameters

A statistical model is said to be identifiable if there exists a unique set of parameter values that can explain the observed data. Mathematically, a statistical model is identifiable if the following conditions hold:
Let $\mathcal{X}$ be the sample space and $f_\theta(x)$ be the probability density function or probability mass function of the statistical model, indexed by the parameter vector $\theta \in \Theta$.

1. θ is finite-dimensional and $\Theta \subseteq \mathbb{R}^d$ for some integer d.
2. For any $\theta_1, \theta_2 \in \Theta$, if $f_{\theta_1}(x) = f_{\theta_2}(x)$ for all $x \in \mathcal{X}$, then $\theta_1 = \theta_2$. In other words, the likelihood function, which is defined as $L(\theta; x) = f_\theta(x)$, is injective with respect to θ.

Condition 1 ensures that the parameter vector is finite-dimensional and belongs to some subset of Euclidean space. Condition 2 ensures that there exists a one-to-one correspondence between the parameter values and the probability distribution, which is necessary for unique estimation of the parameters. These one-to-one correspondence in the context of Fed-ComBat in Equation (3) ensures that there the true expectation effect is captured by $\phi(\mathbf{x}_{ij}; \boldsymbol{\theta}) + \alpha_g$ and not by $\hat{\gamma_{ig}}$. Equivalently for Equation (4). In practice this is translated to:

- No allowing intercepts in the neural network architectures so only one is estimated by α_g.
- Using a design matrix $\mathbf{z}$ shown in Equation (A.1) that allow the estimation of a single set of intercepts that can then be decomposed $\boldsymbol{\beta}_g = \{\alpha_g + \gamma_{ig}\}_{i=1}^{S}$.

S4. CRediT author statement

Santiago Silva: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Visualization. **Neil Oxtoby:** Resources, Data Curation, Writing - Review & Editing. **Andre Altmann:** Supervision, Validation, Writing - Review & Editing. **Marco Lorenzi:** Supervision, Methodology, Validation, Writing - Review & Editing.