Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data

Santiago Silva[1], Boris A. Gutman[2], Eduardo Romero[3], Paul M. Thompson[4], Andre Altmann[5], Marco Lorenzi[1], and for ADNI, PPMI, and UK Biobank[*]

[1] *Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France*
[2] *Department of Biomedical Engineering, Illinois Institute of Technology, USA*
[3] *CIM@LAB, Universidad Nacional de Colombia, Bogotá, Colombia*
[4] *USC Stevens Institute for Neuroimaging and Informatics, Los Angeles, USA*
[5] *Centre for Medical Image Computing, UCL, London, UK*

Abstract

At this moment, databanks worldwide contain brain images of previously unimaginable numbers. Combined with developments in data science, these massive data provide the potential to better understand the genetic underpinnings of brain diseases. However, different datasets, which are stored at different institutions, cannot always be shared directly due to privacy and legal concerns, thus limiting the full exploitation of big data in the study of brain disorders. Here we propose a *federated learning* framework for securely accessing and meta-analyzing any biomedical data without sharing individual information. We illustrate our framework by investigating brain structural relationships across diseases and clinical cohorts. The framework is first tested on synthetic data and then applied to multi-centric, multi-database studies including ADNI, PPMI, MIRIAD and UK Biobank, showing the potential of the approach for further applications in distributed analysis of multi-centric cohorts.

Keywords: Federated learning, distributed databases, PCA, SVD, meta-analysis, brain disease.

1 Introduction

Nowadays, a large amount of magnetic resonance images (MRI) scans are stored across a vast number of clinical centers and institutions. Researchers are currently analyzing these large datasets to understand the underpinnings of brain diseases. However, due to privacy concerns and legal complexities, data hosted in different centers cannot always be directly shared. In practice, data sharing is also hampered by the need to transfer large volumes of biomedical data with the associated bureaucratic burden. This situation led researchers to look for an analysis solution within *meta-analysis* or *federated learning* paradigms. In the federated setting, a model is fitted without sharing individual information across centres, but only model parameters. Meta-analysis instead performs statistical testing by combining results from several independent assays[1], for example by sharing p-values, effect sizes, and/or standard errors across centers.

One of the best examples of such a research approach is the *Enhancing NeuroImaging Genetics through Meta-Analysis* (ENIGMA) consortium (enigma.usc.edu). With a large number of institutions worldwide [2], ENIGMA has become one of the largest networks bringing together multiple groups analyzing neuroimaging data from over 10,000 subjects. However, most of ENIGMA's secure meta-analytic studies in neuroimaging are performed using mass-univariate models.

The main drawback of mass-univariate analysis is that they can only model a single dependent variable at a time. This is a limiting assumption in most of the biomedical scenarios (e.g., neighboring voxels or genetic variations are highly correlated). To overcome this problem, multivariate analysis methods have been proposed to better account for covariance in high-dimensional data.

In a federated analysis context, a few works proposed generalization of standard neuroimaging multivariate analysis methods, such as Independent Component Analysis [3], sparse regression, and parametric statistical testing [4, 5]. Since these methods are mostly based on stochastic gradient descent, a large-number of communications across centers may be required to reach convergence. Therefore, there is a risk of computational and practical bottlenecks when applied to multicentric high-dimensional data.

Lorenzi *et* al.[6, 7] proposed a multivariate dimensionality reduction approach based on eigen-value decomposition. This approach does not require iteration over centers, and was demonstrated on the analysis of the joint variability in imaging-genetics data. However, this framework is still of limited practical utility in real applications, as data harmonization (e.g., standardization and covariate adjustment) should be also consistently performed in a federated way.

Herein we contribute to the state-of-the-art in federated analysis of neuroimaging data by proposing an end-to-end framework for data standardization, confounding factors correction, and multivariate analysis of variability of high-dimensional features. To avoid the potential bottlenecks of gradient-based optimization, the framework is based

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, 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: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Also, from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. This research has been conducted using the UK Biobank Resource.

on schemes analysis through *Alternating Direction Method of Multipliers* (ADMM) reducing the amount of iterations.

We illustrate the framework leveraging on the ENIMGA Shape tool, to provide a first application of federated analysis compatible with the standard ENIGMA pipelines. It should be noted that, even though this work is here illustrated for the analysis of subcortical brain changes in neurological diseases, it can be extended to general multimodal multivariate analysis, such as to imaging-genetics studies.

The framework is benchmarked on synthetic data (section 3.1). It is then applied to the analysis of subcortical thickness and shape features across diseases from multi-centric, multi-database data including: Alzheimer's disease (AD), progressive and non-progressive mild cognitive impairment (MCIc, MCInc), Parkinson's disease (PD) and healthy individuals (HC) (section 3.2).

2 Methods

Biomedical data is assumed to be partitioned across different centers restricting the access to individual information. However, centers can individually share model parameters and run pipelines for feature extraction.

We denote the *global* data (e.g., image arrays) and covariates (e.g., age, sex information) as respectively $\mathbf{X}$ and $\mathbf{Y}$, obtained by concatenating respectively data and covariates of each center. Although these data matrices cannot be computed in practice, this notation will be used to illustrate the proposed methodology. In the global setting, variability analysis can be performed by analyzing the *global data covariance matrix* $\mathbf{S}$.

For each center $c \in \{1, \ldots, C\}$ with N_c subjects each, we denote by $\mathbf{X}_c = (\mathbf{x}_i)_{i=1}^{N_c}$ and $\mathbf{Y}_c = (\mathbf{y}_i)_{i=1}^{N_c}$ the *local* data and covariates. The feature-wise mean and standard deviation vectors of each center are denoted as $\bar{\mathbf{x}}_c$ and σ_c.

The proposed framework is illustrated in Figure 1 and discussed in section 2.1. It is based on three main steps: 1) data standardization, 2) correction from confounding factors and 3) variability analysis.

Data standardization is a data pre-processing step, aiming to enhance the stability of the analysis and easing the comparison across features. In practice, each feature is mapped to the same space by centering data feature-wise to zero-mean and by scaling to unit standard deviation. However, this is ideally performed with respect to the statistics from the whole study (*global statistics*). This issue is addressed by proposing a distributed standardization method in section 2.1.1.

Confounding factors have a *biasing* effect on the data. To correct for this bias, it is usually assumed a linear effect of the confounders $\widehat{\mathbf{X}} = \mathbf{YW}$, that must be estimated and removed. However, for a distributed scenario, computing $\mathbf{W}$ is not straightforward, since the global data matrix cannot be computed. We propose in section 2.1.2 to use *Alternating Direction Method of Multipliers* (ADMM) to estimate a matrix $\widetilde{\mathbf{W}}$ shared among centers, closely approximating $\mathbf{W}$. In particular, we show that $\widetilde{\mathbf{W}}$ can be estimated in a federated way, without sharing local data $\mathbf{X}_c$ nor covariates $\mathbf{Y}_c$.

Finally, through federated principal component analysis (fPCA), we obtain a low dimensional representation of the full data without ever sharing any center's individual information $\mathbf{X}_c, \mathbf{Y}_c$ (section 2.1.3).

2.1 Federated Analysis Framework

2.1.1 Standardization

The mean and standard deviation vectors can be initialized to $\bar{\mathrm{x}}_0 = 0$ and $\bar{\sigma}_0 = 0$. They can be iteratively updated with the information of each center by following standard forms [8], by simply transmitting the quantities $\bar{\mathrm{x}}_c$ and σ_c from center to center. For each center the scaled data is denoted as $\widehat{\mathbf{X}}_c$ and keeps the dimensions of $\mathbf{X}_c$.

2.1.2 Correction from confounding factors

Under the assumption of a linear relationship between data and confounders, the parameters matrix $\mathbf{W}$ can be estimated via *ordinary least squares*, through the minimization of the error function $f(\mathbf{W}) = \left\|\mathbf{Y} - \widehat{\mathbf{X}}\mathbf{W}\right\|^2$.

In a distributed setting, this approach can be performed locally in each center, ultimately leading to C independent solutions. However, this would introduce a bias in the correction, as covariates are accounted for differently across centers.

To solve this issue, we propose to constrain the local solutions to a global one shared across centers. In this way, the subsequent correction can be consistently performed with respect to the estimated global parameters. Thus, we can formulate the problem of constrained regression via ADMM [9].

For a given error function $f_c(\mathbf{W}_c) = \left\|\mathbf{Y}_c - \widehat{\mathbf{X}}_c\mathbf{W}_c\right\|^2$ associated with each center c and constrained to a estimated global matrix of weights $\widetilde{\mathbf{W}}$ we can pose:

$$\text{minimize} \sum_{c=1}^{C} f_c(\mathbf{W}_c), \quad \text{subject to}\, \mathbf{W}_c = \widetilde{\mathbf{W}}, \quad \forall c.$$

As this is a constrained minimization problem, the extended Lagrangian can be calculated as a combination of the parameters from each center (eqn. 1).

$$L_\rho(\mathbf{W}, \widetilde{\mathbf{W}}, \alpha) = \sum_{c=1}^{C} \Big(f_c(\mathbf{W}_c) + \langle \alpha_c, \mathbf{W}_c - \widetilde{\mathbf{W}} \rangle + \frac{\rho}{2} \left\|\mathbf{W}_c - \widetilde{\mathbf{W}}\right\|_2^2 \Big) \quad (1)$$

Where ρ is a penalty factor (or *dual update step length*) regulating the minimization step length for $\mathbf{W}$ and $\widetilde{\mathbf{W}}$. α is a *dual variable* to decouple the optimization of $\mathbf{W}$ and $\widetilde{\mathbf{W}}$.

Optimization is performed as follows: i) Each center independently calculates the local parameters $\mathbf{W}_c$ and α_c (eqn. 2 and 3); ii) the parameters $\mathbf{W}_c$ and α_c are shared to estimate the global parameters $\widetilde{\mathbf{W}}$ (eqn. 4). We note that this last step is performed without sharing either local data or covariates. The parameters $\widetilde{\mathbf{W}}$ are subsequently re-transmitted to the centers and the whole procedure is iterated until convergence:

$$\begin{aligned}\mathbf{W}_c^{(k+1)} &:= \arg\min_{\mathbf{W}_c} L_\rho(\mathbf{W}_c, \widetilde{\mathbf{W}}^{(k)}, \alpha_c^{(k)}) \\ &= \left(\widehat{\mathbf{X}}_c'\widehat{\mathbf{X}}_c + \frac{\rho}{2}\mathbf{I}\right)^{-1}\left(\widehat{\mathbf{X}}_c'\mathbf{Y}_c - \frac{1}{2}\alpha_c^{(k)} + \frac{\rho}{2}\tilde{\mathbf{W}}_c^{(k)}\right) \quad (2)\end{aligned}$$

$$\alpha_c^{(k+1)} := \alpha_c^{(k)} + \rho\big(\mathbf{W}_c^{(k+1)} - \widetilde{\mathbf{W}}^{(k+1)}\big) \quad (3)$$

$$\widetilde{\mathbf{W}}^{(k+1)} := \arg\min_{\widetilde{\mathbf{W}}} L_\rho(\mathbf{W}_c^{(k+1)}, \widetilde{\mathbf{W}}, \alpha_c^{(k)}) = \frac{1}{C}\sum_c^C \left(\frac{\alpha_c^{(k)}}{\rho} + \mathbf{W}_c^{(k+1)}\right) \quad (4)$$

After convergence, $\widetilde{\mathbf{W}}$ is shared across centers, and used to consistently account for covariates by subtracting their effect from the structural data to obtain the corrected observation matrix: $\mathbf{E}_c = \widehat{\mathbf{X}}_c - \mathbf{Y}_c\widetilde{\mathbf{W}}$.

2.1.3 Federated PCA (fPCA)

Principal components analysis (PCA) is a standard approach for dimensionality reduction assuming that the largest amount of information is contained in the directions $\mathbf{U}$ (components) of greater variability. Data can be thus represented by projecting on the low-dimensional space spanned by the main components: $\widehat{\mathbf{E}} = \mathbf{E}\mathbf{U}$.

From the eigen-value decomposition of the global covariance matrix $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}'$, the first m-eigen-modes $\mathbf{U} = (\mathbf{u}_j)_{j=1}^m$ provide a low-dimensional representation of the overall variation in $\mathbf{E}$. In our federated setting, we note that $\mathbf{S}$ is the algebraic sum of the *local covariance matrices* $\mathbf{S} = \mathbf{E}\mathbf{E}' = \sum_{c=1}^S \mathbf{E}_c\mathbf{E}_c'$. Based on this observation, Lorenzi *et* al. proposed to share only the eigen-modes and values of the covariance matrix of each center avoiding the access to individual data [6]. However, sharing the local-covariance-matrices can still be prohibitive as the dimension is $(N_{\text{features}} \times N_{\text{features}})$. For this reason, it was proposed to further reduce the dimensionality of the problem by sharing only the principal eigen-components associated with the local covariance matrices: $\mathbf{S} \approx \sum_{c=1}^C \mathbf{U}_c\mathbf{\Sigma}_c^2\mathbf{U}_c'$. From the practical point of view, computing the eigen-components can be efficiently performed by solving the eigen-problem associated with the matrix $(\mathbf{X}_c\mathbf{X}_c')^2$ which is usually of much smaller dimension $(N_c \times N_c)$ [10].

In what follows, the number of components shared across centers is automatically defined by fixing a threshold of 80% on the associated *explained variability* contained in $\mathbf{\Sigma}_c$.

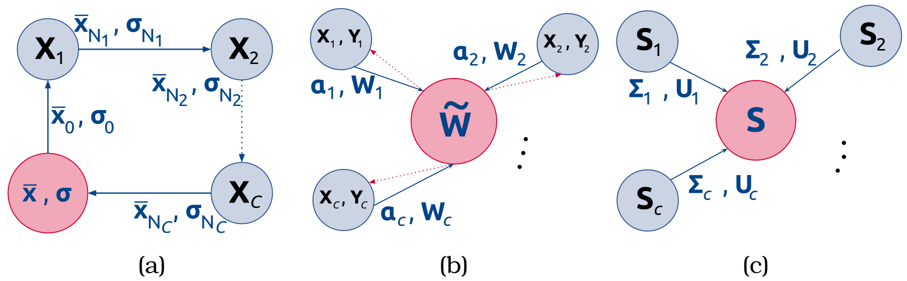

Figure 1: Data flow to obtain: (a) the global statistics $\bar{x}$ and σ, (b) the shared parameter matrix $\widehat{\mathbf{W}}$ to correct from covariates and (c) the approximated global covariance matrix S. Red node: master; blue nodes: local centers. Arrows denote the data flows from centers (blue) and from the master (red).

Database (total)	**ADNI (802)**				**MIRIAD (68)**		**PPMI (232)**	**UK Biobank (208)**
Group	HC	MCInc	MCIc	AD	HC	AD	PD	HC
N (females)	109 (115)	62 (119)	78 (130)	89 (100)	11 (12)	26 (19)	85 (147)	116 (92)
Age ± sd	75.79 (4.99)	74.93 (7.72)	74.54 (7.09)	75.19 (7.48)	69 (7.18)	69.17 (7.06)	60.69 (8.95)	60.72 (7.52)

Table 1: Data used in this study. Each study here represents an independent center. The centers are jointly analyzed through the federated analysis proposed in Section 2.1.

3 Experiments

3.1 Synthetic Data

We randomly generated $\mathbf{Y}$ and $\mathbf{W}$ matrices. The data matrix was subsequently computed as $\mathbf{X} = \mathbf{YW}$, and corrupted with Gaussian noise $\mathcal{N}(0, \sigma)$, with σ set to 20% of $\|\mathbf{X}\|$. Then, $\mathbf{X}$ and $\mathbf{Y}$ were split in C centers of equal sample size. Our federated framework was then applied for each scenario across 200 folds, and convergence analyzed as shown in Figure 2.

3.2 Real Data: Neuroimaging

Data. T1-weighted MRI scans at baseline were analyzed from several research databases (table 1). In total, we included data for 455 controls (HC), 181 with non-progressive MCI (MCInc), 208 progressive (MCIc), 234 Alzheimer's disease (AD), 232 with Parkinson's disease (PD).

Feature extraction. ENIGMA Shape Analysis was applied to the MRI data of each center [11, 12]. In our analysis we extracted: a) radial distance (an approximate measure of thickness) and, b) the $\log$ of the Jacobian determinant (surface area dilation/contraction) for each vertex of the following subcortical regions: hippocampi, amygdalae, thalami, pallidum, caudate nuclei, putamen and accumbens nuclei. The overall data dimension is of 54,240 features.

Federated analysis. Each database of table 1 was modeled as an independent center. Sex, Age and Age^2 were used to correct the vertex-

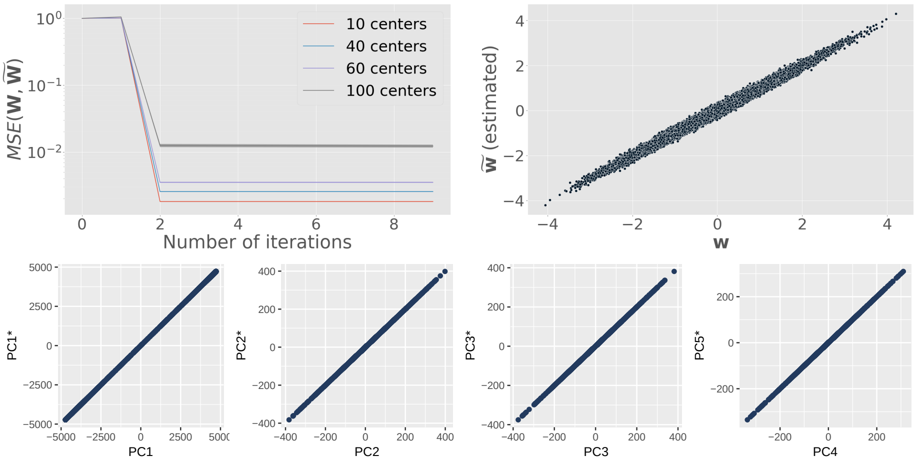

Figure 2: **Top-left:** Mean square error (MSE) between $\mathbf{W}$ and $\widetilde{\mathbf{W}}$ for different numbers of centers. $N = 2400$, $N_{\text{features}} = 50,000$ and $\dim(\mathbf{y}) = 20$. **Top-right:** Single-column of $\mathbf{W}$ vs $\widetilde{\mathbf{W}}$ for $C = 100$. **Bottom:** Principal components (PC) vs federated ones (PC*) for 100 centers.

wise shape data according to 2.1.1 and 2.1.2. For ADMM, convergence was ensured through 10 iterations. Finally, the analysis of the variability was performed according to 2.1.3.

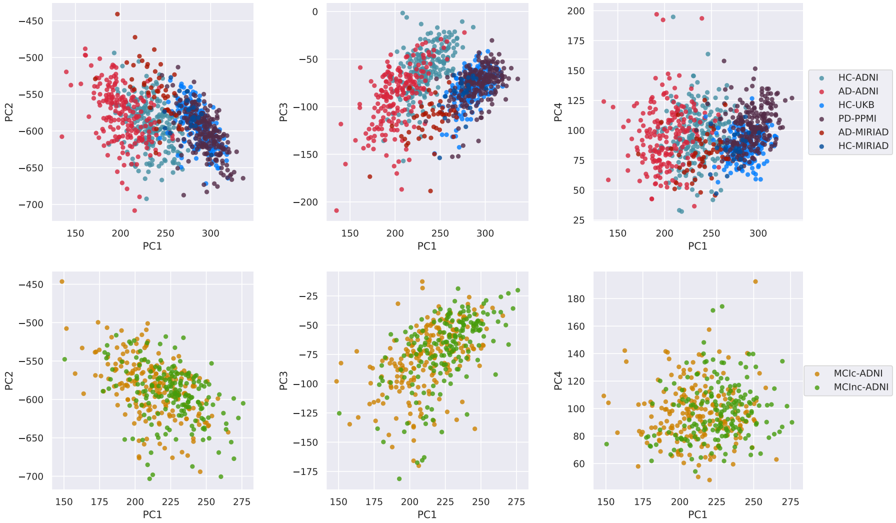

Figure 3: Data projected on the first 4 components. AD vs controls from different centers (top). MCI progressive and stable from ADNI (bottom). Federated PCA was performed on the whole data obtained from the 4 centers (table 1).

Results. The projection in the latent space spanned by the federated principal components is shown in Figure 3. To ease visualization, the

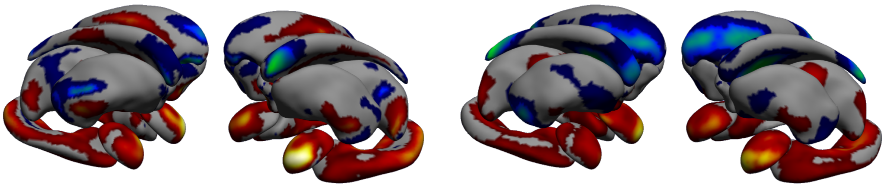

Figure 4: First principal component estimated with the proposed federated framework. The component maps prevalently hippocampi and amigdalae. **Left:** Thickness. **Right:** Log-Jacobians.

projection for MCI converters and those who remained stable is shown in the bottom panel. Figure 4 shows the weight maps associated to the first principal component. We note that principal components 1 to 3 identify a variability from healthy to AD consistent across centers. Moreover, healthy ADNI participants are in between the AD subjects and the rest of the population. This result may denote some residual effect of Age on the resulting imaging features, even after correction. Interestingly, the issue of "leaking" spurious variability of confounders after correction has been already reported in a number of multi-centric studies, and is matter of ongoing research [13, 14]. Finally we note that PD subjects are generally similar to the healthy individuals with respect to the modelled subcortical information.

4 Conclusions

In this work we proposed, tested, and validated a fully consistent framework for federated analysis of distributed biomedical data. Further developments of this study will extend the proposed analysis to large-scale imaging genetics data, such as in the context of the ENIGMA meta-study.

5 Acknowledgments

This work was supported by the French government, through the UCA-JEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01 (project Meta- ImaGen). A.A. holds an Medical Research Council eMedLab Medical Bioinformatics Career Development Fellowship. This work was supported by the Medical Research Council [grant number MR/L016311/1]. Additional support is from NIH grants RF1AG051710, R01MH116147 and R56AG058854. The full list of acknowledgments for the data providers is available at the supplementary material in section 6.

References

[1] Junfeng Sun et al. Meta-analysis of Clinical Trials. *Principles and Practice of Clinical Research*, pages 317–327, jan 2018.

[2] Paul M. Thompson et al. The ENIGMA Consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior*, 8(2):153–182, 2014.

[3] Bradley T Baker et al. Large scale collaboration with autonomy: Decentralized data ica. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015.

[4] Jing Ming et al. Coinstac: Decentralizing the future of brain imaging analysis. *F1000Research*, 6, 2017.

[5] Sergey M. Plis et al. Coinstac: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Frontiers in Neuroscience*, 10:365, 2016.

[6] Marco Lorenzi et al. Secure multivariate large-scale multi-centric analysis through on-line learning: an imaging genetics case study. In *12th International Symposium on Medical Information Processing and Analysis*, volume 10160, page 1016016. International Society for Optics and Photonics, 2017.

[7] Marco Lorenzi et al. 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(12):3162–3167, 3 2018.

[8] B. P. Welford. Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics*, 4(3):419, 8 1962.

[9] Stephen Boyd. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2010.

[10] Keith J. Worsley et al. Comparing functional connectivity via thresholding correlations and singular value decomposition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):913–920, 2005.

[11] Benjamin S.C. Wade et al. Mapping abnormal subcortical brain morphometry in an elderly HIV + cohort. *NeuroImage: Clinical*, 9:564–573, 2015.

[12] Gennady V. Roshchupkin et al. Heritability of the shape of subcortical brain structures in the general population. *Nature Communications*, 7:1–8, 2016.

[13] Jacob Westfall and Tal Yarkoni. Statistically controlling for confounding constructs is harder than you think. *PloS one*, 11(3):e0152719, 2016.

[14] Stephen M Smith and Thomas E Nichols. Statistical challenges in "Big Data" human neuroimaging. *Neuron*, 97(2):263–268, 2018.

6 Supplementary material: Acknowledgements

6.1 Funding

This work was supported by the French government, through the UCA-JEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01 (project Meta- ImaGen). A.A. holds an Medical Research Council eMedLab Medical Bioinformatics Career Development Fellowship. This work was supported by the Medical Research Council [grant number MR/L016311/1]. Additional support is from NIH grants RF1AG051710, R01MH116147 and R56AG058854.

6.2 The Alzheimer's Disease Neuroimaging Initiative (ADNI)

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, 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: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

6.3 The Parkinson's Progression Markers Initiative (PPMI)

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org.

PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/fundingpartners.

6.4 UK Biobank

This research has been conducted using the UK Biobank Resource. Additional information can be found at: https://www.ukbiobank.ac.uk.