Removal of Scanner Effects in Covariance Improves Multivariate Pattern Analysis in Neuroimaging Data

Andrew A. Chen[a,b,*], Joanne C. Beer[a], Nicholas J. Tustison[c], Philip A. Cook[d], Russell T. Shinohara[a,b,1], Haochang Shou[a,b,1], and the Alzheimer's Disease Neuroimaging Initiative[e]

[a]Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA 19104

[b]Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA 19104

[c]Department of Radiology and Medical Imaging, University of Virginia, Irvine, Charlottesville, VA 22903

[d]Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104

[e]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

[1]Equal contribution

***Correspondence: Andrew A. Chen**
andrewac@pennmedicine.upenn.edu

Abstract

To acquire larger samples for answering complex questions in neuroscience, researchers have increasingly turned to multi-site neuroimaging studies. However, these studies are hindered by differences in images acquired across multiple scanners. These effects have been shown to bias comparison between scanners, mask biologically meaningful associations, and even introduce spurious associations. To address this, the field has focused on harmonizing data by removing scanner-related effects in the mean and variance of measurements. Contemporaneously with the increase in popularity of multi-center imaging, the use of multivariate pattern analysis has also become commonplace. These approaches have been shown to provide improved sensitivity, specificity, and power due to their modeling the joint relationship across measurements in the brain. In this work, we demonstrate that the currently available methods for removing scanner effects are inherently insufficient for MVPA. This stems from the fact that no currently available harmonization approach has addressed how correlations between measurements can vary across scanners. Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) is used to show that considerable differences in covariance exist across scanners and that the state-of-the-art harmonization techniques do not address this issue. We also propose a novel methodology that harmonizes covariance of multivariate image measurements across scanners and demonstrate its improved performance in data harmonization, which further facilitates more power for detection of clinically relevant associations.

1 Introduction

The need for larger samples in human subjects research have led to a growing number of multi-site studies that aggregate data across multiple locations. This trend is especially prevalent in neuroimaging research where the reliability and generalizabilty of findings from the conventional single-site studies are often limited by the ability to recruit and study sufficiently large and representative samples from the population. Many consortia have been formed to address such issues (Mueller et al., 2005; Sudlow et al., 2015; Trivedi et al., 2016; Van Essen et al., 2013). The larger samples obtained through these efforts promote greater power to detect significant associations as well as better generalizability of results. However, these study designs also introduce heterogeneity in acquisition and processing that, if not appropriately addressed, may impact study findings.

Several researchers have determined that variability driven by scanner, often called scanner effects, reduce the reliability of derived measurements and can introduce bias. Neu-

roimaging measurements have been repeatedly shown to be affected by scanner manufacturer, model, magnetic field strength, head coil, voxel size, and acquisition parameters (Han et al., 2006; Kruggel et al., 2010; Reig et al., 2009; Wonderlick et al., 2009). Yet even in scanners of the exact same model and manufacturer, differences still exist for certain neuroimaging biomarkers (Takao et al., 2011).

Until recently, neuroimaging analyses primarily involved mass univariate testing which treats features as independent. Under this paradigm, the impact of scanner effects is through changes in the mean and variance of measurements. Increasingly, researchers have used sets of features as patterns for prediction algorithms in a framework called multivariate pattern analysis (MVPA). This approach has become a powerful tool in diverse research topics including pain perception (Smith et al., 2017), neural representations (Haxby et al., 2014), and psychiatric illnesses (Koutsouleris et al., 2014). One of the major benefits of MVPA is that it leverages the joint distribution and correlation structure among multivariate brain features in order to better characterize a phenotype of interest (O'Toole et al., 2007). As a result, scanner effects on the covariance of measurements are likely to impact findings substantially. In fact, a recent investigation showed that MVPA was able to detect scanner with high accuracy and that the detection of sex depended heavily on the scanners included in the training and test data (Glocker et al., 2019).

The major statistical harmonization techniques employed in neuroimaging have generally corrected for differences across scanners in mean and variance, but not covariance (Fortin et al., 2016, 2018; Rao et al., 2017; Yamashita et al., 2019). Increasingly, the ComBat model (Johnson et al., 2007) has become a popular harmonization technique in neuroimaging and has been successfully applied to structural and functional measures (Bartlett et al., 2018; Fortin et al., 2017, 2018; Marek et al., 2019; Yu et al., 2018). However, this model does not address potential scanner effects in covariance. Recently, another stream of data-driven harmonization methods have aimed to apply machine learning algorithms such as generative adversarial network (GAN) or distance-based methods to unify distributions of measurements across scanners, but these methods shift the original data distributions inexplicitly and have not been tested for their potential influence on MVPA (Nguyen et al., 2018; Zhou et al., 2018).

In this paper, we examine whether scanner effects influence MVPA results. In particular, we study the cortical thickness measurements derived from images acquired by the Alzheimer's Disease Neuroimaging Initiative (ADNI) and demonstrate the existence of scanner effects in covariance of structural imaging measures. We then propose a novel harmonization method called Correcting Covariance Batch Effects (CovBat) that removes scanner effects in mean, variance, and covariance. We apply CovBat and show that within-scanner

correlation matrices are successfully harmonized. Furthermore, we find that machine learning methods are unable to distinguish scanner manufacturer after our proposed harmonization is applied, and that the CovBat-harmonized data facilitate more accurate prediction of disease group. We also assess the performance of the proposed method in simulated data, and again find that the method mitigates scanner effects and improves detection of meaningful associations. Our results demonstrate the need to consider covariance in harmonization methods, and suggest a novel procedure that can be applied to better harmonize data from multi-site imaging studies.

2 Methods

2.1 ADNI data analysis

All data for this paper are obtained from ADNI (http://adni.loni.usc.edu/ and processed using the ANTs longitudinal single-subject template pipeline (Tustison et al., 2019) with code available on GitHub (https://github.com/ntustison/CrossLong).

We briefly summarize the steps involved. First, we download raw MP-RAGE or the equivalent sequence for GE scanners from the ADNI-1 database. For each subject, we estimate a single-subject template from all the image timepoints. After rigid spatial normalization to this single-subject template, each normalized timepoint image is then processed using the single image cortical thickness pipeline consisting of brain extraction, denoising, N4 bias correction, Atropos n-tissue segmentation, and registration-based cortical thickness estimation. For our analyses, we use the 62 cortical thickness values from the baseline scans. The correlation matrices for these cortical thicknesses in the three largest sites are shown in Fig. 1.

We define scanner based on information contained within the Digital Imaging and Communications in Medicine (DICOM) files for each scan. Specifically, subjects are considered to be acquired on the same scanner if they share the same location of scan, scanner manufacturer, scanner model, head coil, and magnetic field strength. In total, this definition yields 142 distinct scanners of which 78 had less than three subjects and were removed from analyses. The final sample consists of 505 subjects across 64 scanners, with 213 subjects imaged on scanners manufactured by Siemens, 70 by Philips, and 222 by GE. The sample has a mean age of 75.3 (SD 6.70) and is comprised of 278 (55%) males, 115 (22.8%) Alzheimer's disease (AD) patients, 239 (47.3%) late mild cognitive impairment (LMCI), and 151 (29.9%) cognitively normal (CN) individuals.

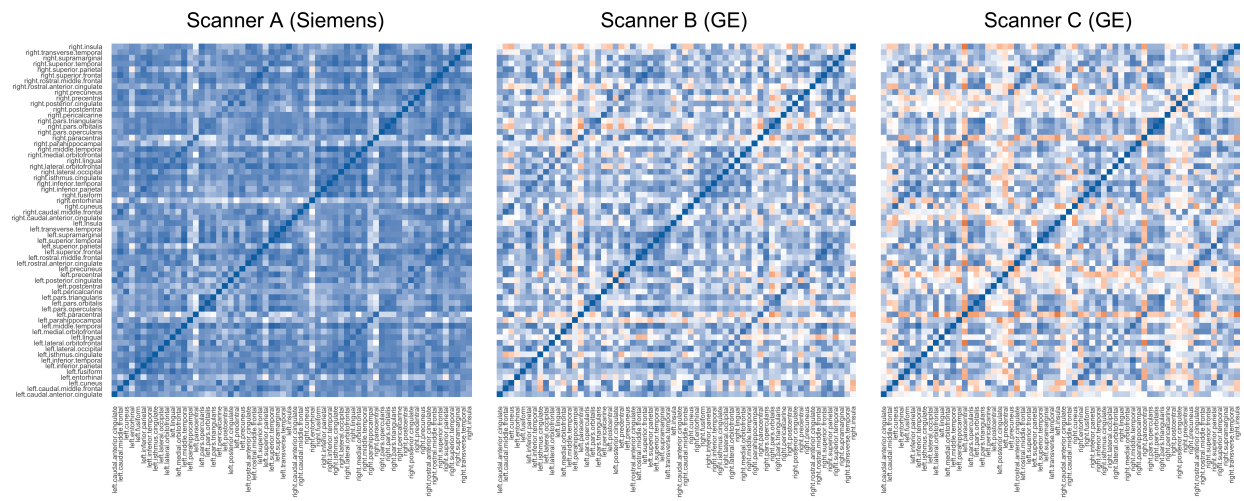

Figure 1: **Correlation matrices for cortical thickness values acquired on three scanners.** Scanner A was a Siemens Symphony 1.5T scanner and the other scanners are General Electric Signa Excite 1.5T scanners.

2.2 Combatting Batch Effects

To introduce notation, we first review ComBat (Fortin et al., 2017, 2018; Johnson et al., 2007) for harmonization of neuroimaging measures. ComBat seeks to remove the mean and variance scanner effects in the data in an empirical Bayes framework. Let $\boldsymbol{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijp})^T,\ i = 1, 2, \ldots, M,\ j = 1, 2, \ldots, n_i$ denote the $p \times 1$ vectors of observed data where p is the number of features. Our goal is to harmonize these vectors across the M scanners indexed by i. ComBat assumes that the features indexed by v follow

$$y_{ijv} = \alpha_v + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_v + \gamma_{iv} + \delta_{iv} e_{ijv} \tag{1}$$

where α_{ijv} is the intercept, $\boldsymbol{x}_{ij}$ is the vector of covariates, $\boldsymbol{\beta}_v$ is the vector of regression coefficients, γ_{iv} is the mean scanner effect, and δ_{iv} is the variance scanner effect. The errors e_{ijv} are assumed to follow $e_{ijv} \sim N(0, \sigma_v^2)$. ComBat finds empirical Bayes point estimates γ_{iv}^* and δ_{iv}^* by pooling information across features then residualizes with respect to these estimates to obtain ComBat-adjusted data

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \hat{\alpha}_v - \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \hat{\alpha}_v + \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_v \tag{2}$$

2.3 Correcting Covariance Batch Effects

We propose the CovBat algorithm, which accounts for the joint distribution of ComBat-adjusted observations as follows:

Step 1. We first perform ComBat to remove the mean and variance shifts in the marginal distributions of the cortical thickness measures. Then, we additionally residualize with respect to the intercept and covariates to obtain ComBat-adjusted residuals denoted $\boldsymbol{e}_{ij}^{ComBat} = (e_{ij1}^{ComBat}, e_{ij2}^{ComBat}, \ldots, e_{ijp}^{ComBat})^T$ where

$$e_{ijv}^{ComBat} = \frac{y_{ijv} - \hat{\alpha}_v - \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_v - \gamma_{iv}^*}{\delta_{iv}^*} \tag{3}$$

Step 2. The $\boldsymbol{e}_{ij}^{ComBat}$ are assumed to have mean 0; their covariance matrices which we denote by Σ_i, however, may differ across scanners. We thus perform principal components analysis (PCA) on the full data residuals and represent the full-data covariance matrix as $\Sigma = \sum_{k=1}^{q} \lambda_k \boldsymbol{\phi}_k \boldsymbol{\phi}_k^T$ where $q = \min(\sum_{i=1}^{M} n_i, p)$. The ComBat-adjusted residuals can then be expressed as $e_{ij}^{ComBat} = \sum_{k=1}^{q} \xi_{ijk} \boldsymbol{\phi}_k$, where the λ_k are the eigenvalues of Σ, $\boldsymbol{\phi}_k$ are the principal components obtained as the eigenvectors of Σ and ξ_{ijk} are the principal component scores. After applying this decomposition, the scanner-specific covariance matrices can be expressed using the full data eigenvectors as $\Sigma_i = \sum_{k=1}^{q} \lambda_{ik} \boldsymbol{\phi}_k \boldsymbol{\phi}_k^T$ where λ_{ik} are scanner-specific eigenvalues. This model assumes that the covariance scanner effect is contained within the λ_{ik}, which can be approximated as the sample variance of the principal component scores ξ_{ijk}.

Step 3. Thus, we posit:

$$\xi_{ijk} = \mu_{ik} + \rho_{ik} \epsilon_{ijk} \tag{4}$$

where $\epsilon_{ijk} \sim \mathrm{N}(0, \tau_k^2)$ and μ_{ik}, ρ_{ik} are the center and scale parameters corresponding to principal components $k = 1, 2, \ldots K$ where $K \leq q$ is a tuning parameter chosen to capture the majority of the variation in the observations. Note that this is analogous to the ComBat model, applied to each of the k principal component scores instead of the original measures. We can then estimate each of the K pairs of center and scale parameters by finding the values that bring each scanner's mean and variance in scores to the pooled mean and variance. We then remove the scanner effect in the scores via $\xi_{ijk}^{CovBat} = (\xi_{ijk} - \hat{\mu}_{ik})/\hat{\rho}_{ik}$.

Step 4. We obtain CovBat-adjusted residuals $\boldsymbol{e}_{ij}^{CovBat} = (e_{ij1}^{CovBat}, e_{ij2}^{CovBat}, \ldots, e_{ijp}^{CovBat})^T$ by projecting the adjusted scores back into the residual space via

$$\boldsymbol{e}_{ij}^{CovBat} = \sum_{k=1}^{K} \xi_{ijk}^{CovBat} \boldsymbol{\phi}_k + \sum_{l=K+1}^{q} \xi_{ijl} \boldsymbol{\phi}_l \tag{5}$$

We then add the intercepts and covariates effects estimated in Step 1 to obtain CovBat-adjusted observations

$$y_{ijv}^{CovBat} = e_{ijv}^{CovBat} + \hat{\alpha}_v + \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_v \tag{6}$$

2.4 Multivariate pattern analysis experiments

We evaluate whether the harmonization procedures affect the results of MVPA using the neuroimaging measures as patterns. We achieve this through a Monte Carlo split-sample experiment where we i) randomly split the subjects into 50% training set and 50% validation set, ii) train a random forests algorithm to detect either scanner manufacturer or a binary clinical covariate, and iii) assess predictive performance on the validation set via AUC. We train separate models for unharmonized, ComBat-harmonized, and CovBat-harmonized data where both harmonization methods are performed including age, sex, and diagnosis status as covariates. We perform steps (i)-(iii) 100 times for each dataset and report the mean AUC along with standard deviation. For these experiments, lower AUC for detection of scanner manufacturer and higher AUC for detection of clinical covariates would indicate improved harmonization. For prediction of scanner manufacturer, we avoid the possibility that scanner could be detected through the covariates age, sex, and disease status by residualizing out these variables from each cortical thickness value via linear models.

2.5 Simulation design

Let $\boldsymbol{y}_{ij}$, $i = 1, 2, 3$, $j = 1, 2, \ldots, 250$ be vectors of length 62 representing simulated cortical thickness values for three sites, each with 250 observations. The $\boldsymbol{y}_{ij}$ are generated using the following model:

$$\boldsymbol{y}_{ij} = \boldsymbol{\alpha} + x_{ij}\boldsymbol{\beta} + \boldsymbol{\gamma}_i + \boldsymbol{\delta}_i^T \boldsymbol{e}_{ij}$$

where $\boldsymbol{\alpha}$ is the sample mean vector of Scanner B observations in the ADNI data, $\boldsymbol{\beta}$ is the vector of predictor effects on the mean, and x_{ij} is a binary covariate drawn from a Bernoulli distribution with probability 0.25. The mean site effects $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{i62})^T$ are vectors drawn from independent and identically distributed (i.i.d.) normal distributions with mean zero and standard deviation 0.1. The variance site effects $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{i62})^T$ are vectors drawn from site-specific inverse gamma distributions with chosen parameters. For our simulations, we chose to distinguish the site-specific scaling factors by assuming $\delta_{1v} \overset{i.i.d}{\sim}$ Inverse Gamma $(46, 50)$, $\delta_{2v} \overset{i.i.d}{\sim}$ Inverse Gamma $(51, 50)$, and $\delta_{3v} \overset{i.i.d}{\sim}$ Inverse Gamma $(56, 50)$ for $v = 1, 2, \ldots, 62$. The error terms $\boldsymbol{e}_{ij}$ are generated via

$$\boldsymbol{e}_{ij} \sim N(\mathbf{0}, D_{ij}\Sigma_{ij}D_{ij})$$

where $\Sigma_{ij} = S + x_{ij}\Psi + \Omega_i$, S is the sample correlation matrix of Scanner B observations in the ADNI data, Ψ is a chosen predictor-driven covariance shift matrix, and Ω_i are site-specific covariance shift matrices. The matrices $D_{ij} = \mathbf{diag}(\boldsymbol{d}_{ij})$ where $d_{ijk} = 1/\sqrt{(\Sigma_{ij})_{kk}}$

for $k = 1, 2, \ldots, 62$ ensure that these covariance effects do not modify the marginal variances of $\boldsymbol{e}_{ij}$. To constrain the covariance matrices to be positive definite, we set the negative eigenvalues of Σ_{ij} equal to a small constant, 10^{-12}.

We generate 1000 datasets using the above model and perform experiments to test these hypotheses under a similar paradigm as the MVPA experiments implemented on the ADNI dataset. For each dataset, we i) randomly split the sample into 50% training data and 50% testing data, ii) train a random forests algorithm to recognize if the observations either are from Site 1 or have the simulated predictor, and iii) assess predictive performance on the testing data. Lower AUC for detection of Site 1 and higher AUC for detection of the simulated predictor indicate improved harmonization. For prediction of Site 1, we avoid the possibility that site could be detected through the predictor by using linear models to residualize out the predictor from each simulated cortical thickness value.

2.5.1 Predictor without effect on covariance

In the ComBat simulation, we impose site effects in mean and variance while having the predictor affect only the mean of the observations. We choose $\beta_i = -0.5$ for 15 regions of interest in both the left and right hemispheres to impose that about half of the ROIs are negatively associated with the predictor. We also choose the Ω_i and Ψ to be 62×62 zero matrices to ensure that the covariance does not depend on site or the predictor.

In the Predictor Affects Mean simulation, we again impose that the predictor only affects the mean of the measurements but also introduce a site effect in covariance. We keep the same $\boldsymbol{\beta}$ and Ψ as in the ComBat simulation. However, we choose Ω_i to be distinct 62×62 high-rank matrices to distinguish covariance structures across sites.

2.5.2 Predictor with effect on covariance

In the Predictor Affects Covariance simulation, we assume that the predictor affects not only mean, but also variance and covariance. We choose the predictor effect on covariance to be proportional to a site's covariance shift. This scenario represents a situation where detection of the predictor using MVPA would be highly influenced by the presence of site effects. We use the same $\boldsymbol{\beta}$ value as in the ComBat and Predictor Affects Mean simulations but choose Ψ to be related to Ω_2 to force confounding of Site 3 and predictor effects on covariance. To achieve this, we set $\Psi = -(3/4)\Omega_3$.

In the Covariance Only simulation, we assume that both site and the predictor influence the covariance, not the mean or variance. We fix $\boldsymbol{\gamma} = \mathbf{0}$ and $\boldsymbol{\delta} = \mathbf{1}$ to remove site effects in mean and variance while also using the same Ω_i as in the Predictor Affects Mean and

	A (Siemens)	B (General Electric)	C (General Electric)	p
Number of Subjects	23	20	20	
Age (mean (SD))	74.48 (5.13)	76.90 (8.18)	78.78 (6.19)	0.11
Diagnosis (%)				0.57
AD	7 (30.4)	6 (30.0)	5 (25.0)	
CN	6 (26.1)	5 (25.0)	2 (10.0)	
LMCI	10 (43.5)	9 (45.0)	13 (65.0)	
Male (%)	10 (43.5)	13 (65.0)	16 (80.0)	0.05

Table 1: **ADNI demographics by scanner for the three scanners with the largest number of acquired subjects.** Manufacturer of each scanner is displayed in parentheses. ANOVA p-values for testing differences in the mean of continuous variables and Chi-squared test p-values for testing the differences in categorical variables are reported in the rightmost column.

Predictor Affects Covariance simulations. Furthermore, we modify the predictor effect by setting $\boldsymbol{\beta} = \mathbf{0}$ while keeping $\Psi = -(3/4)\Omega_3$ for the predictor effect in covariance.

3 Results

3.1 CovBat reduces covariance scanner effect

We first apply CovBat to observations acquired on the three scanners with the largest number of subjects. Scanner A consists of 23 subjects acquired on a Siemens Symphony 1.5T scanner while scanners B and C each consist of 20 subjects acquired on GE Signa Excite 1.5T scanners. See Table 1 for demographic details. We observe that demographic variables differ across scanner so we residualize each cortical thickness measure on age, sex, and diagnosis status to obtain the correlation structure independent of these clinical covariates. Fig. 2 shows the correlation matrices for each scanner using the residualized cortical measures both before and after CovBat. The differences between the unharmonized correlation matrices are striking. Especially notable are the increased positive correlations across most pairs of cortical regions in Scanner A and the weakened right-left correlations in Scanner C visible as the diagonal line in the top-left and bottom-right quadrants. Visually, the correlation structures are considerably more similar across scanners after CovBat; the correlation structures of Scanners A and B are almost indistinguishable after this adjustment.

We also compare with harmonization via ComBat, and report our quantitative results

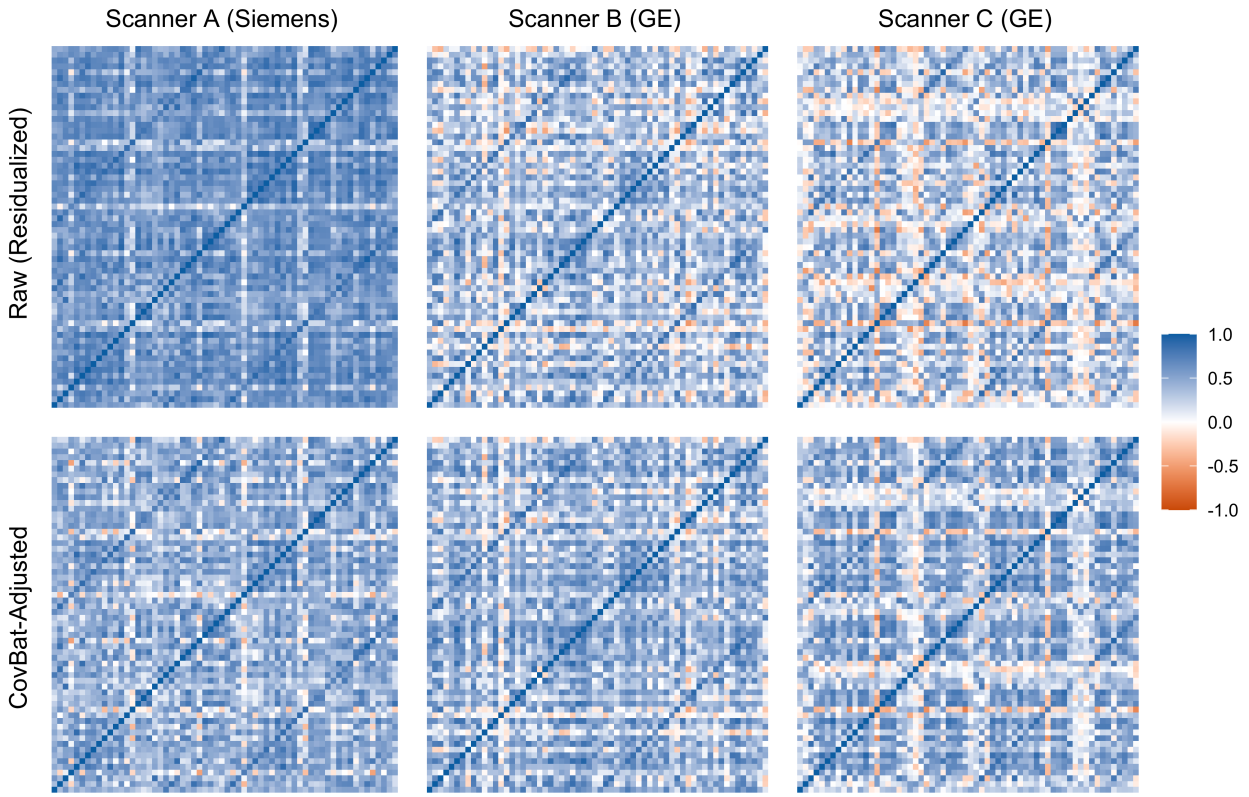

Figure 2: **Correlation matrices for cortical thickness values acquired on three scanners before and after CovBat harmonization.** Scanner A is a Siemens Symphony 1.5T scanner and the other scanners are General Electric Signa Excite 1.5T scanners. ComBat-adjusted matrices are visually indistinguishable from unharmonized.

for ComBat-adjusted as well as CovBat-adjusted correlation matrices in Table 2. A tuning parameter of the CovBat model is the desired proportion of variance explained in the dimension reduction space, which we selected at 80% (26 PCs). To ensure that our results do not depend strongly on the choice of tuning parameter, we also report the minimum and maximum of the pairwise Frobenius norms after applying CovBat with percent variation explained ranging from 50% (10 PCs) to 99% (53 PCs). We report the results of this sensitivity analysis in parentheses. We find that ComBat adjustment does not harmonize the correlation matrices whereas CovBat adjustment shows large reductions in the between-scanner distances across a range of tuning parameter choices.

3.2 CovBat impairs detection of scanner

To evaluate the potential impact of scanner effects in covariance using MVPA, we conduct a Monte-Carlo split-sample experiment for prediction of scanner manufacturer labels using

	Unharmonized	ComBat	CovBat
A,B	22.52	22.52	14.85 (14.78 - 16.18)
A,C	27.58	27.58	16.37 (16.35 - 18.19)
B,C	17.57	17.57	16.05 (16.02 - 16.44)

Table 2: **Pairwise distances between scanner-specific correlation matrices.** Differences in correlation structure between scanner are reported via the Frobenius distance between correlation matrices calculated across observations acquired on each scanner. Results from adjusting the number of PC scores ranging from those explaining 56.2% to 99.9% of variation are shown in parentheses as the minimum and maximum pairwise Frobenius norms across the range.

data before and after harmonization with existing methods. We train using data harmonized with the state-of-the-art ComBat method and our proposed method, CovBat. Fig. 3a shows that Siemens scanners are easily identifiable based on unharmonized cortical thickness measurements (area-under-the-curve (AUC) 0.89 ± 0.02), which is consistent with recent findings (Glocker et al., 2019). We also note that scanner manufacturer is still detected after ComBat is applied (AUC 0.66 ± 0.03). After CovBat, the machine learning method's performance for differentiating between scanners is close to chance (AUC 0.54 ± 0.03).

3.3 CovBat recovers biological associations

It is well-known that cortical thickness differs substantially by sex and Alzheimer's disease status (Lerch et al., 2005; Sowell et al., 2007). To assess whether CovBat maintains biological associations of interest, we perform two MVPA experiments using random forests to classify healthy versus Alzheimer's disease (AD) and to differentiate patients by sex. Fig. 3b,c show that detection of these biological differences is considerably improved by either harmonization method, but the proposed CovBat approach shows an even greater performance improvement. For detection of AD, the mean AUC increases from 0.73 ($\pm$0.03) in raw data to 0.78 ($\pm$0.03) in ComBat-harmonized data to 0.79 ($\pm$0.03) in CovBat-harmonized data. Similarly, the mean AUC for detection of sex increased from 0.66 ($\pm$0.03) to 0.69 ($\pm$0.03) to 0.70 ($\pm$0.03). These findings suggest that CovBat not only provides thorough removal of scanner effects, but also helps to recover clinical associations. **Appendix A.2** shows that these results for both detection of site and biological associations largely hold even when CovBat is trained on a subset of the data.

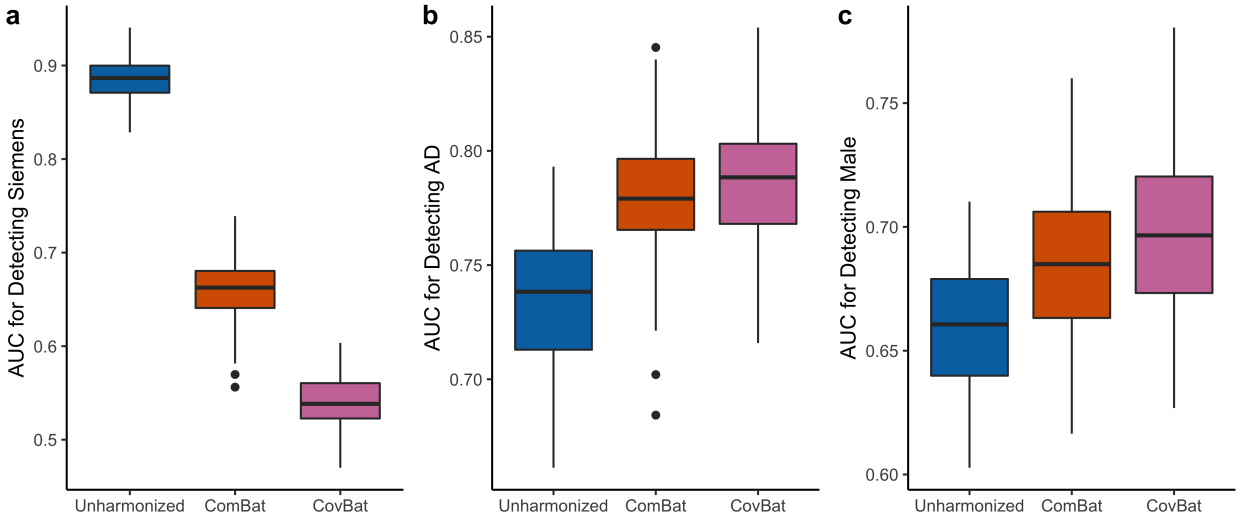

Figure 3: **Multivariate pattern analysis experiments for detection of scanner manufacturer, sex, and Alzheimer's disease status using cortical thickness data.** The data is randomly split into 50% training and 50% validation then used to train a random forests algorithm to predict a specified trait. AUC values from 100 repetitions of this analysis are reported for unharmonized, ComBat-adjusted, and CovBat-adjusted data. **a**, boxplot showing results for detecting if subjects were acquired on a Siemens scanner. Results for detection of Alzheimer's disease status are shown in **b** and results for detection of sex are shown in **c**.

3.4 Simulation studies

We then evaluate CovBat in simulated datasets that consist of observations from three sites with distinct mean and covariance structures (Methods). We begin by examining a ComBat simulation, where we generate data from the original ComBat model. That is, we impose mean and variance site effects while also simulating a predictor that has an effect on the mean. Fig. 4a,b show that CovBat performs almost identically to ComBat in this scenario, showing that our method performs competitively in the absence of covariance effects. In our Predictor Affects Mean simulation, we introduce a covariance site effect and show in Fig. 4c,d that CovBat substantially reduces the chance of detecting site and performs similarly to ComBat for detection of the predictor. These simulations demonstrate that CovBat performs at least as well or better than ComBat when the predictor affects only the mean of the observations.

We then incorporate a predictor effect on covariance in our Predictor Affects Covariance simulation and show that CovBat outperforms ComBat in harmonizing covariance (see Ta-

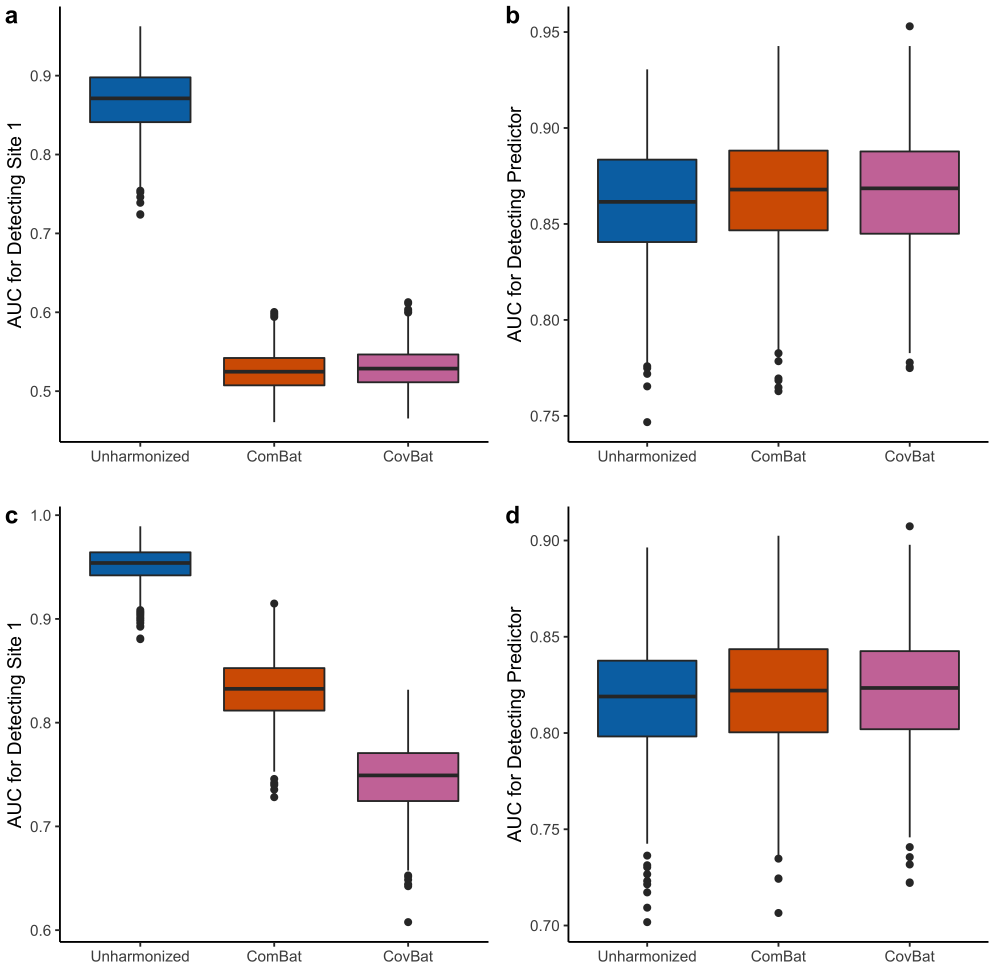

Figure 4: **Results from MVPA simulations for detection of site and for detection of a binary predictor in the absence of predictor effects on covariance.** For each of 1000 simulations, the data is randomly split into 50% training and 50% validation then a random forests algorithm is trained using the training set to predict either Site 1 or the presence of the binary predictor. **a**, Boxplot showing Site 1 detection in the ComBat simulation with high accuracy in the unharmonized data (AUC 0.87 ± 0.04) and low accuracy after both ComBat (0.53 ± 0.02) and CovBat (0.53 ± 0.03). **b**, Boxplot showing predictor detection in the ComBat simulation where mean AUC for detection of the predictor using the unharmonized data is the lowest (0.86 ± 0.03), ComBat shows some performance improvement, (0.87 ± 0.03), and CovBat performs similarly (0.86 ± 0.03). **c**, Boxplot showing Site 1 detection in the Predictor Affects Mean simulation with high accuracy in unharmonized data (AUC 0.95 ± 0.02), some reduction after ComBat (0.83 ± 0.03) and substantial reduction after CovBat (0.74 ± 0.03) **d**, Boxplot showing predictor detection in the Predictor Affects Mean simulation with lowest mean AUC using the unharmonized data (0.82 ± 0.03) and similar accuracy after both ComBat (0.82 ± 0.03) and CovBat (0.82 ± 0.03).

	Unharmonized	ComBat	CovBat
1,2	17.11 (1.19)	17.11 (1.19)	11.77 (0.48)
1,3	16.57 (1.22)	16.57 (1.22)	11.63 (0.52)
2,3	10.67 (0.42)	10.67 (0.42)	11.17 (0.37)

Table 3: **Mean and standard deviation of pairwise Frobenius norms over 1000 simulations.** Standard deviations are reported in parentheses. These norms reflect the difference between site-specific correlation matrices for the Predictor Confounded with Site simulation. Results show that ComBat does not harmonize these correlation matrices whereas CovBat substantially reduces between-scanner distances involving Site 1.

ble 3), reducing detection of site (Fig. 5a), and recovering the association with the predictor (Fig. 5b). In order to further emphasize the importance of covariance effects, we investigate a Covariance Only simulation where both site and predictor effects exist only in the covariance of observations, but not the mean or variance. In unharmonized data, we observe high mean AUC values for detection of Site 1 and detection of the predictor, both of which are essentially unaffected after implementing ComBat. After CovBat though, we see substantial improvements on both metrics (Fig. 5c,d). While CovBat performs well in both simulations, we find that CovBat does not entirely remove the severe covariance site effect. Regardless, we observe that CovBat offers notable improvements over ComBat owing to its harmonization of covariance.

4 Discussion

The growing number of multi-site studies across diverse fields has spurred the development of harmonization methods that are general, but also account for field-specific challenges. In neuroimaging research, the rise of MVPA has established an unmet need for harmonization of covariance. We demonstrated that strong scanner effects in covariance exist and could influence downstream MVPA experiments, which remain after performing the state-of-the-art harmonization. We then proposed a novel method demonstrated to be effective in removing scanner differences in covariance and improving the detection of biological associations via MVPA. Simulation studies replicate these observations and suggest that improvement in predictor detection is linked to a predictor effect on covariance, even in the absence of mean and variance effects. This finding suggests that future work could investigate how predictors influence covariance and identify situations that particularly benefit from CovBat harmonization. While our study focused on structural neuroimaging data, our findings ex-

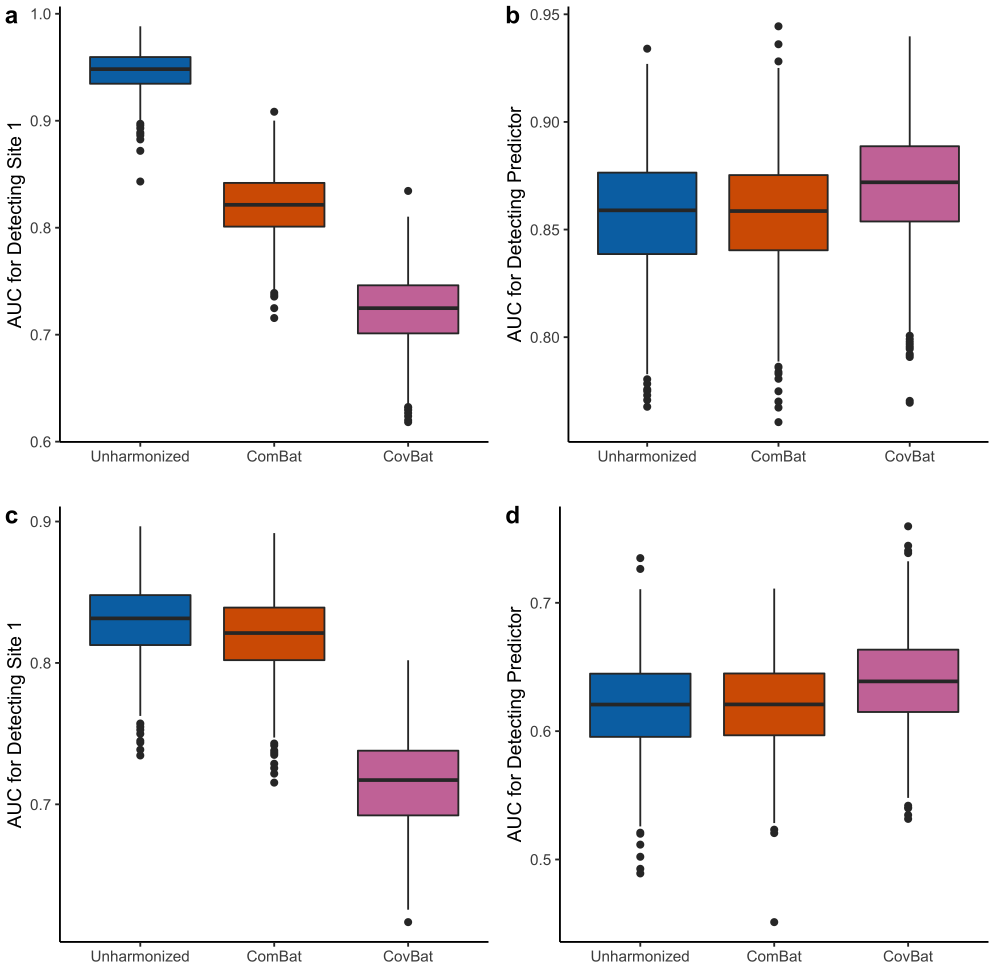

Figure 5: **Results from MVPA simulations for detection of site and for detection of a binary predictor where the predictor has effects on covariance.** For each of 1000 simulations, the data is randomly split into 50% training and 50% validation then a random forests algorithm is trained using the training set to predict either Site 1 or the presence of the binary predictor. **a**, Boxplot showing Site 1 detection in the Predictor Affects Covariance simulation with high accuracy in the unharmonized data (AUC 0.95 ± 0.02), lower accuracy after ComBat (0.82 ± 0.03), and lowest accuracy after CovBat (0.72 ± 0.03). **b**, Boxplot showing predictor detection in the ComBat simulation where mean AUC for detection of the predictor is similar in both the unharmonized data (0.86 ± 0.03) and ComBat-adjusted data (0.86 ± 0.03), but higher in the CovBat-adjusted data (0.87 ± 0.03). **c**, Boxplot showing Site 1 detection in the Covariance Only simulation with high accuracy in unharmonized data (AUC 0.81 ± 0.03), almost no reduction after ComBat (0.80 ± 0.03), and substantial reduction after CovBat (0.69 ± 0.03) **d**, Boxplot showing predictor detection in the Covariance Only simulation with lower accuracy using the unharmonized data (0.69 ± 0.04), similar accuracy after ComBat (0.69 ± 0.04), and notably higher accuracy after CovBat (0.69 ± 0.04).

tend directly to functional, metabolic, and other imaging modalities. Further studies should also determine the extent to which multivariate statistical and machine learning studies of genomic data are susceptible to the biases documented.

5 Software

All of the postprocessing analysis was performed in the R statistical software (V3.6.1). CovBat is available for both R and Python (https://github.com/andy1764/CovBat_Harmonization). Reference implementations for ComBat are available in R and Matlab (https://github.com/Jfortin1/ComBatHarmonization) and in Python (https://github.com/ncullen93/neuroCombat).

6 Acknowledgements

This work was supported by the National Institute of Neurological Disorders and Stroke (grant numbers R01 NS085211 and R01 NS060910), the National Multiple Sclerosis Society (RG-1707-28586) and a seed grant from the University of Pennsylvania Center for Biomedical Image Computing and Analytics (CBICA). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

The majority of the data used in this paper are derived from the ADNI study. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National

Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Competing interests

The authors declare no competing interests.

Author contributions

R.T.S. and H.S. contributed equally to this work. **Andrew A. Chen:** Formal analysis, Conceptualization, Methodology, Software, Investigation, Writing - original Draft **Joanne C. Beer:** Data curation, Writing - review & editing **Nicholas J. Tustison:** Data curation, Writing - review & editing **Philip A. Cook:** Data curation, Writing - review & editing **Russell T. Shinohara:** Supervision, Conceptualization, Methodology, Writing - review & editing **Haochang Shou:** Supervision, Conceptualization, Methodology, Writing - review & editing

Appendix A.

A.1 CovBat improves prediction of age

ComBat has previously been shown to improve age prediction via linear regression and support vector regression (Fortin et al., 2018). To evaluate whether this result holds in ADNI data for ComBat or CovBat, we propose an additional MVPA experiment. We i) randomly split the subjects into 50% training set and 50% validation set, ii) train a random forests algorithm to predict age, and iii) assess predictive performance on the validation set via root-mean-square-error (RMSE). We perform steps (i)-(iii) 100 times each for unharmonized, ComBat-adjusted, and CovBat-adjusted data. Fig. A.1. shows that the mean RMSE for age prediction decreases from 6.05 ($\pm$0.22) in unharmonized to 5.81 ($\pm$0.23) in ComBat-adjusted data to 5.75 ($\pm$0.21) in CovBat-adjusted data. These findings are consistent with previous work (Fortin et al., 2018) and show that CovBat provides additional recovery of the association between cortical thickness and age.

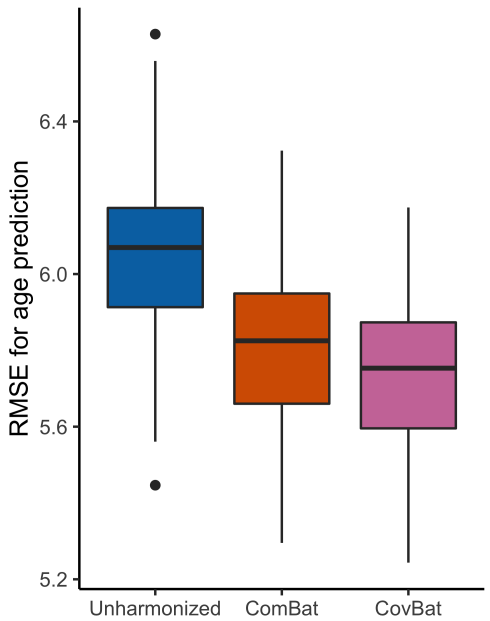

Figure A.1.: **Multivariate pattern analysis experiments for detection of age using cortical thickness data.** The data is randomly split into 50% training and 50% validation then used to train a random forests algorithm to predict age. RMSE values from 100 repetitions of this analysis are reported for unharmonized, ComBat-adjusted, and CovBat-adjusted data.

A.2 Parameter estimation using training subset

Both ComBat and CovBat estimate and residualize out the covariate effects using the full data; however, there are cases were only a subset of the data is available when performing harmonization. For instance, if a group of subjects has already been acquired, prediction on subjects subsequently acquired on the same scanners could only leverage data from the original sample. In this scenario, the new sample can be harmonized using ComBat or CovBat by estimating the covariate effect using the original sample, then proceeding with subsequent steps as usual.

We evaluate this modification by repeating our main MVPA analyses using ADNI data with different subsampling of the patients. Specifically, we replace step (i) in our MVPA experiments by instead splitting the sample into 270 training subjects and 235 testing subjects such that both the train and test sets contain at least one subject acquired on each scanner. We then apply ComBat and CovBat by estimating the covariate effects using only the training subjects. Fig. A.2. shows the results for all MVPA experiments. Detection of site (AUC 0.89 ± 0.02 in raw data) still worsens after ComBat (0.66 ± 0.03) and is almost at

chance after CovBat (0.54 ± 0.03). For detection of AD, improvements over unharmonized (AUC 0.74 ± 0.03) are still demonstrated after ComBat adjustment (0.77 ± 0.03) and CovBat adjustment (0.78 ± 0.02). For detection of male, lesser improvement is observed from unharmonized (AUC 0.67 ± 0.03) to ComBat (0.68 ± 0.03) to CovBat (0.68 ± 0.03). Mean RMSE for age prediction decreases from 5.99 ($\pm$0.22) in unharmonized to 5.86 ($\pm$0.23) in ComBat-adjusted data to 5.82 ($\pm$0.22) in CovBat-adjusted data. Overall, the results appear quite similar to harmonization using the full dataset, showing that CovBat performs well even when only a limited training subset is available.

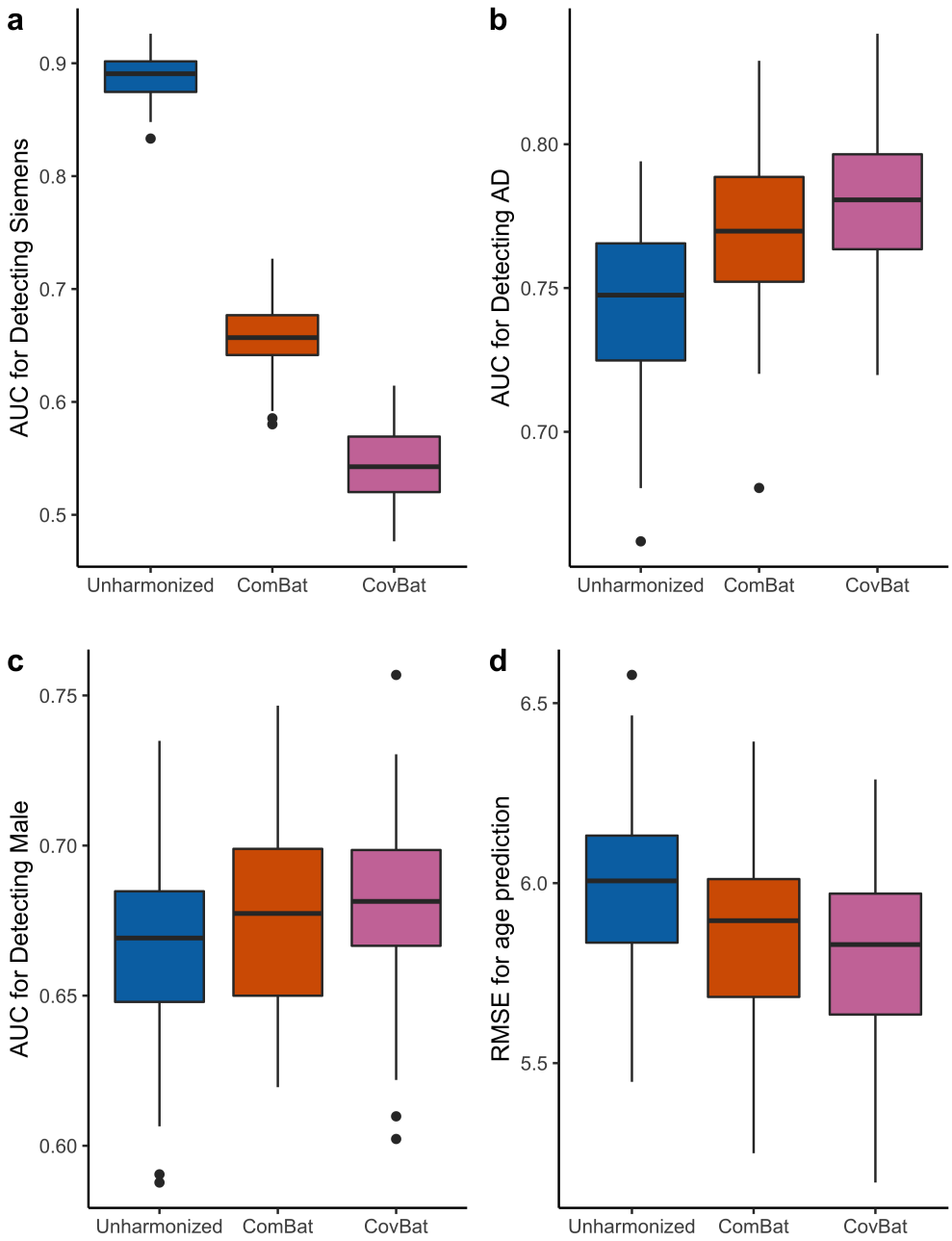

Figure A.2.: **MVPA experiment results for harmonization using only training data.** The data is randomly split into 270 training subjects and 235 testing subjects such that every site is represented in each group. The training set is then used to train a random forests algorithm to predict scanner or patient characteristics. **a** shows the AUC values for detection of site. AUC values for detection of AD are displayed in **b** and detection of male in **c**. RMSE values for prediction of age are displayed in **d**.

References

Bartlett, E. A., DeLorenzo, C., Sharma, P., Yang, J., Zhang, M., Petkova, E., Weissman, M., McGrath, P. J., Fava, M., Ogden, R. T., Kurian, B. T., Malchow, A., Cooper, C. M., Trombello, J. M., McInnis, M., Adams, P., Oquendo, M. A., Pizzagalli, D. A., Trivedi, M. & Parsey, R. V. (2018). Pretreatment and early-treatment cortical thickness is associated with SSRI treatment response in major depressive disorder. *Neuropsychopharmacology* **43**, 2221–2230.

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., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M. & Shinohara, R. T. (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., Schultz, R. T., Verma, R. & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161**, 149–170.

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

Glocker, B., Robinson, R., Castro, D. C., Dou, Q. & Konukoglu, E. (2019). Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects. *arXiv:1910.04597 [cs, eess, q-bio]* .

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B. & Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage* **32**, 180–194.

Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience* **37**, 435–456.

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

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., Pantelis, C. & Meisenzahl, E. (2014). Accelerated Brain Aging in Schizophrenia and Beyond: A Neuroanatomical Marker of Psychiatric Disorders. *Schizophrenia Bulletin* **40**, 1140–1153.

Kruggel, F., Turner, J., Muftuler, L. T. & Alzheimer's Disease Neuroimaging Initiative (2010). Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. *NeuroImage* **49**, 2123–2133.

Lerch, J. P., Pruessner, J. C., Zijdenbos, A., Hampel, H., Teipel, S. J. & Evans, A. C. (2005). Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. *Cerebral Cortex (New York, N.Y.: 1991)* **15**, 995–1001.

Marek, S., Tervo-Clemmens, B., Nielsen, A. N., Wheelock, M. D., Miller, R. L., Laumann, T. O., Earl, E., Foran, W. W., Cordova, M., Doyle, O., Perrone, A., Miranda-Dominguez, O., Feczko, E., Sturgeon, D., Graham, A., Hermosillo, R., Snider, K., Galassi, A., Nagel, B. J., Ewing, S. W. F., Eggebrecht, A. T., Garavan, H., Dale, A. M., Greene, D. J., Barch, D. M., Fair, D. A., Luna, B. & Dosenbach, N. U. F. (2019). Identifying reproducible individual differences in childhood functional brain networks: An ABCD study. *Developmental Cognitive Neuroscience* **40**, 100706.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W. & Beckett, L. (2005). The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging clinics of North America* **15**, 869–xii.

Nguyen, H., Morris, R. W., Harris, A. W., Korgoankar, M. S. & Ramos, F. (2018). Correcting differences in multi-site neuroimaging data using Generative Adversarial Networks. *arXiv:1803.09375 [cs]* .

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P. & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience* **19**, 1735–1752.

Rao, A., Monteiro, J. M. & Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage* **150**, 23–49.

Reig, S., Sánchez-González, J., Arango, C., Castro, J., González-Pinto, A., Ortuño, F., Crespo-Facorro, B., Bargalló, N. & Desco, M. (2009). Assessment of the increase in variability when combining volumetric data from different scanners. *Human Brain Mapping* **30**, 355–368.

Smith, A., López-Solà, M., McMahon, K., Pedler, A. & Sterling, M. (2017). Multivariate pattern analysis utilizing structural or functional MRI-In individuals with musculoskeletal pain and healthy controls: A systematic review. *Seminars in Arthritis and Rheumatism* **47**, 418–431.

Sowell, E. R., Peterson, B. S., Kan, E., Woods, R. P., Yoshii, J., Bansal, R., Xu, D., Zhu, H., Thompson, P. M. & Toga, A. W. (2007). Sex differences in cortical thickness mapped in 176 healthy individuals between 7 and 87 years of age. *Cerebral Cortex (New York, N.Y.: 1991)* **17**, 1550–1560.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T. & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779.

Takao, H., Hayashi, N. & Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. *Journal of Magnetic Resonance Imaging* **34**, 438–444.

Trivedi, M. H., McGrath, P. J., Fava, M., Parsey, R. V., Kurian, B. T., Phillips, M. L., Oquendo, M. A., Bruder, G., Pizzagalli, D., Toups, M., Cooper, C., Adams, P., Weyandt, S., Morris, D. W., Grannemann, B. D., Ogden, R. T., Buckner, R., McInnis, M., Kraemer, H. C., Petkova, E., Carmody, T. J. & Weissman, M. M. (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. *Journal of Psychiatric Research* **78**, 11–23.

Tustison, N. J., Holbrook, A. J., Avants, B. B., Roberts, J. M., Cook, P. A., Reagh, Z. M., Duda, J. T., Stone, J. R., Gillen, D. L., Yassa, M. A. & Initiative, f. t. A. D. N. (2019). Longitudinal Mapping of Cortical Thickness Measurements: An Alzheimer's Disease Neuroimaging Initiative-Based Evaluation Study. *Journal of Alzheimer's Disease* **71**, 165–183.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E. & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An Overview. *NeuroImage* **80**, 62–79.

Wonderlick, J., Ziegler, D., Hosseini-Varnamkhasti, P., Locascio, J., Bakkour, A., van der Kouwe, A., Triantafyllou, C., Corkin, S. & Dickerson, B. (2009). Reliability of MRI-derived cortical and subcortical morphometric measures: Effects of pulse sequence, voxel geometry, and parallel imaging. *NeuroImage* **44**, 1324–1333.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Tanaka, S. C., Kawato, M., Yamashita, O. & Imamizu, H. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLOS Biology* **17**, e3000042.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T. & Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping* **39**, 4213–4227.

Zhou, H. H., Singh, V., Johnson, S. C., Wahba, G. & Initiative, t. A. D. N. (2018). Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proceedings of the National Academy of Sciences* **115**, 1481–1486.