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

Andrew A. Chen[a,2], Joanne C. Beer[a,b], Nicholas J. Tustison[c], Philip A. Cook[d], Russell T. Shinohara[a,b,d,1], Haochang Shou[a,b,d,1], and the Alzheimer's Disease Neuroimaging Initiative[3]

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

[b]Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

[c]Department of Neurobiology and Behavior, University of California, Irvine, Irvine, CA 92697

[d]Center for Biomedical Image Computing and Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104

[1]R.T.S. and H.S. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: andrewac@pennmedicine.upenn.edu

[3]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 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.

2 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. Neuroimaging 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.

Scanner Identified Despite Existing Harmonization

We first examine a subset of the ADNI dataset to determine if covariance among cortical thickness measurements differs across scanners. Our data were obtained from the ADNI-1 database which included a magnetization-prepared gradient echo (MP-RAGE) sequence from Siemens and Philips scanners and a similar works-in-progress MP-RAGE sequence for General Electric (GE) scanners Jack et al. (2010). All of these scans were collected at a magnetic field strength of 1.5 telsa and there were a total of 90 scanners across 58 collaborating institutions. The T1 images were processed using the ANTs cross-sectional cortical thickness pipeline Tustison et al. (2019). Lastly, the cortical thickness measures were derived from the pre-processed images as the average thickness in 62 cortical regions, 31 in each hemisphere, defined through the Desikan-Killiany Atlas Klein & Tourville (2012). For additional information, see *SI Appendix*.

To investigate the potential impact of scanner differences in covariance using MVPA, we conducted an experiment to predict scanner manufacturer labels using data harmonized with existing methods. In particular, we use a Monte Carlo split-sample experiment including data acquired on any scanner used to image three or more subjects. This sample consists of 505 subjects across 64 scanners, with 213 subjects acquired on scanners manufactured by Siemens, 70 by Philips, and 222 by GE. Using the 62 cortical thickness values as inputs, we i) randomly split the sample into 50% training data and 50% testing data; ii) train a random forests algorithm to recognize if a scanner was manufacturered by Siemens, and iii) assess predictive performance on the testing data. We train it using data harmonized via the existing ComBat method and our proposed method CovBat. For all tests, we accounted for the possibility that scanner could be detected through the covariates age, sex, and disease status by residualizing out those variables. We repeat steps (i)-(iii) 100 times and report the area under the receiver operating characteristic curve (AUC) values. Figure 1 shows that Siemens scanners are identifiable based on unharmonized cortical thickness measurements (AUC 0.89±0.02), which is consistent with recent findings Glocker et al. (2019). We also note that scanner manufacturer is still detectable after ComBat is applied (0.66 ± 0.03). After using the proposed method method however, the performance for distinguishing Siemens scanners is close to random (AUC 0.54 ± 0.03).

Statistical Harmonization of Covariance

Combatting Batch Effects

An increasingly popular method for harmonization of neuroimaging measures is a method called ComBat Fortin et al. (2017, 2018); Johnson et al. (2007). This method seeks to remove the mean and variance scanner effects in the data in an empirical Bayes framework. Let $\boldsymbol{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijp})^T$, $i = 1, \ldots, M$, $j = 1, \ldots, n_i$ 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 then finds empirical Bayes point estimates γ_{iv}^* and δ_{iv}^* then residualizes with respect to these estimates to obtain

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

Correcting Covariance Batch Effects

We propose the CovBat algorithm by accounting 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 and additionally residualize with respect to

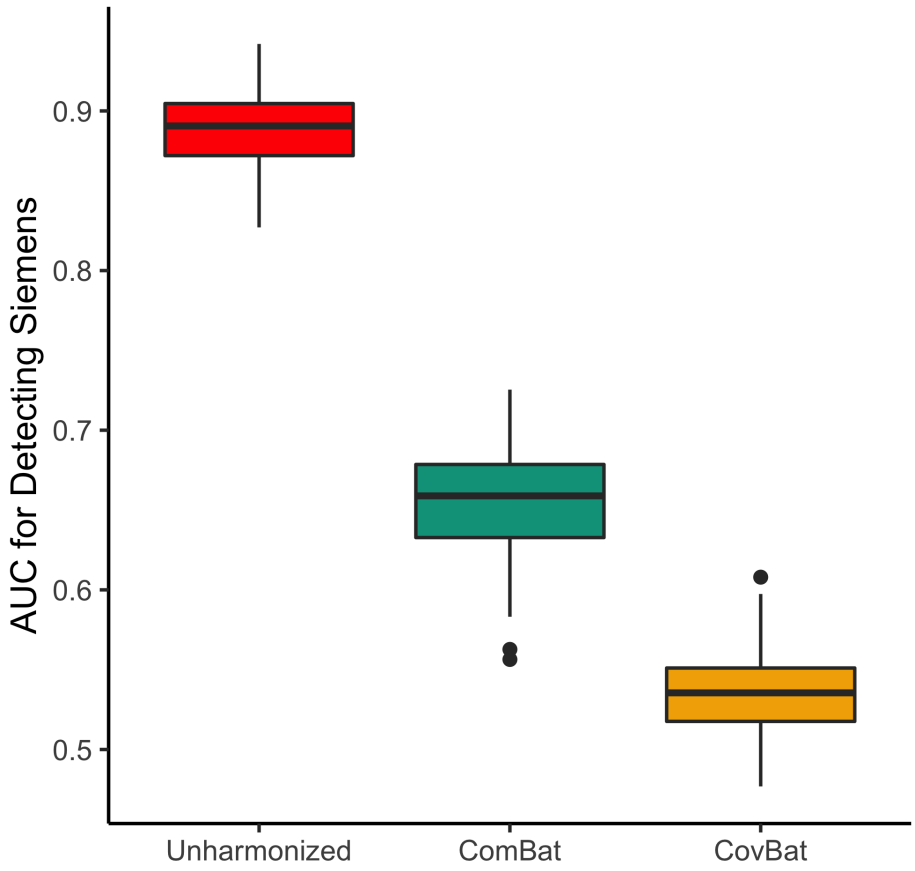

Figure 1: Results of an MVPA experiment for detection of scanner manufacturer using cortical thickness data. The data is randomly split into 50% training and 50% validation then used to train a random forests algorithm to predict if a scanner is manufactured by Siemens. AUC values from 100 repetitions of this analysis are reported for unharmonized, ComBat-adjusted, and CovBat-adjusted data.

the covariates to obtain ComBat-adjusted residuals $\boldsymbol{e}_{ij}^{ComBat} = (e_{ij1}^{ComBat}, e_{ij2}^{ComBat}, \ldots, e_{ijp}^{ComBat})^T$ 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 dataset to obtain 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, ξ_{ijk} are the principal component scores, and $K < q$ is chosen to capture the majority of the variation in the observations. 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 each principal component indexed by k. Note that this is analogous to the ComBat model, applied to each of the k principal component scores instead of the original measures. After imposing an analogous prior on each parameter, 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 $\xi_{ijk}^{CovBat} = \frac{\xi_{ijk} - \hat{\mu}_{ik}}{\hat{\rho}_{ik}}$.

Step 4. Finally, 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

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

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

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

Harmonization Evaluation

We focus our evaluation framework on removal of scanner effects in covariance rather than mean and variance which have been shown to be addressed by ComBat in previous papers Fortin et al. (2017, 2018). Hence, we propose tests that directly assess harmonization of correlation matrices across scanners. Furthermore, we assess the degree to which residual scanner effects can affect comparison between scanners and clinically meaningful associations.

To assess scanner effects in covariance, we examine the correlation matrices before and after harmonization. Additionally, we quantify the similarity of correlation matrices between

scanners by looking at the pairwise Frobenius norms. For this measure, lower values indicate greater harmonization in covariance.

We also evaluate if the harmonization procedures affect the results of MVPA. Similar to the earlier experiments in Section 4 for classifying scanner, we i) randomly split the subjects into 50% training set and 50% validation set; ii) train a random forests algorithm to detect 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 these steps (i)-(iii) 100 times and again report the AUC values. For these experiments, higher AUC would indicate greater ability to recover biologically meaningful associations.

Material and Methods

Here, we provide an overview of the materials and methods used in this paper. More details can be found in *SI Appendix*, and the code for executing the analyses described here is available at (https://github.com/andy1764/CovBat_Harmonization).

ADNI Data

All data for this paper are obtained from ADNI (http://adni.loni.usc.edu/ and processed using the ANTs longitudinal cortical thickness pipeline Tustison et al. (2018) with code available on GitHub (https://github.com/ntustison/CrossLong). We briefly summarize the steps involved. First, raw MP-RAGE or the equivalent sequence for GE scanners are downloaded from the ADNI-1 database. The images are first processed using the ANTs cross-sectional cortical thickness pipeline Tustison et al. (2014), which involves N4 bias correction, brain extraction, Atropos n-tissue segmentation, and registration-based cortical thickness estimation. Then, a single-subject template is created for each individual using all of their repeated scans, and the template is subsequently used in rigid registration of the subject's images. For our analyses, we only use the cortical thickness values of the baseline scans.

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.

Simulation Design

Let $\boldsymbol{y}_{ij}$, $i = 1, 2, 3$, $j = 1, 2, \ldots, 100$ be vectors of length 62 representing the simulated outcome for cortical thickness values in 62 regions. 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 x_{ij} is a single binary covariate drawn from a Bernoulli random variable with probability 0.5, $\boldsymbol{\alpha}$ is the vector of intercepts, $\boldsymbol{\beta}$ is the vector of coefficients, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{i62})^T$ are vectors of region-specific mean shift drawn from independently and identically distributed (i.i.d.) standard normal distributions and $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{i62})^T$ are vectors of region-specific scale shifts drawn from i.i.d. scanner-specific inverse gamma distributions with chosen parameters. For our simulations, we chose to sufficiently distinguish the scanner-specific scaling factors by assuming $\delta_{1v} \sim$ Inverse Gamma$(2, 0.5)$, $\delta_{2v} \sim$ Inverse Gamma$(3, 1)$, and $\delta_{3v} \sim$ Inverse Gamma$(4, 2)$ for $v = 1, 2, \ldots, 62$. The error terms $e_{ij} \sim \mathrm{N}(\mathbf{0}, \Sigma + \Omega_i + x_{ij}\Psi)$ where Σ is the sample covariance matrix of Scanner B in the ADNI analyses, x_{ij} is a single binary covariate, Ω_i are scanner-specific covariance shift matrices, and Ψ is a chosen covariance shift matrix which can be similar to any of the Ω_i. To ensure that the covariance matrices are positive semi-definite, we set the negative eigenvalues equal to a small constant, 10^{-12}. This method for inducing covariance scanner effects ensures flexibility in the direction, complexity, and confounding of the effect. For additional details and results of the specific simulation settings considered in this papers, see *SI Appendix*.

CovBat Reduces Covariance Scanner Effect

We apply CovBat to observations acquired on the three scanners with the largest number of subjects. Scanner A was a Siemens Symphony 1.5T scanner while scanners B and C are GE Signa Excite 1.5T scanners. See *SI Appendix* 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. Figure 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 for ComBat-adjusted as well as CovBat-adjusted correlation matrices in Table 1. 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.

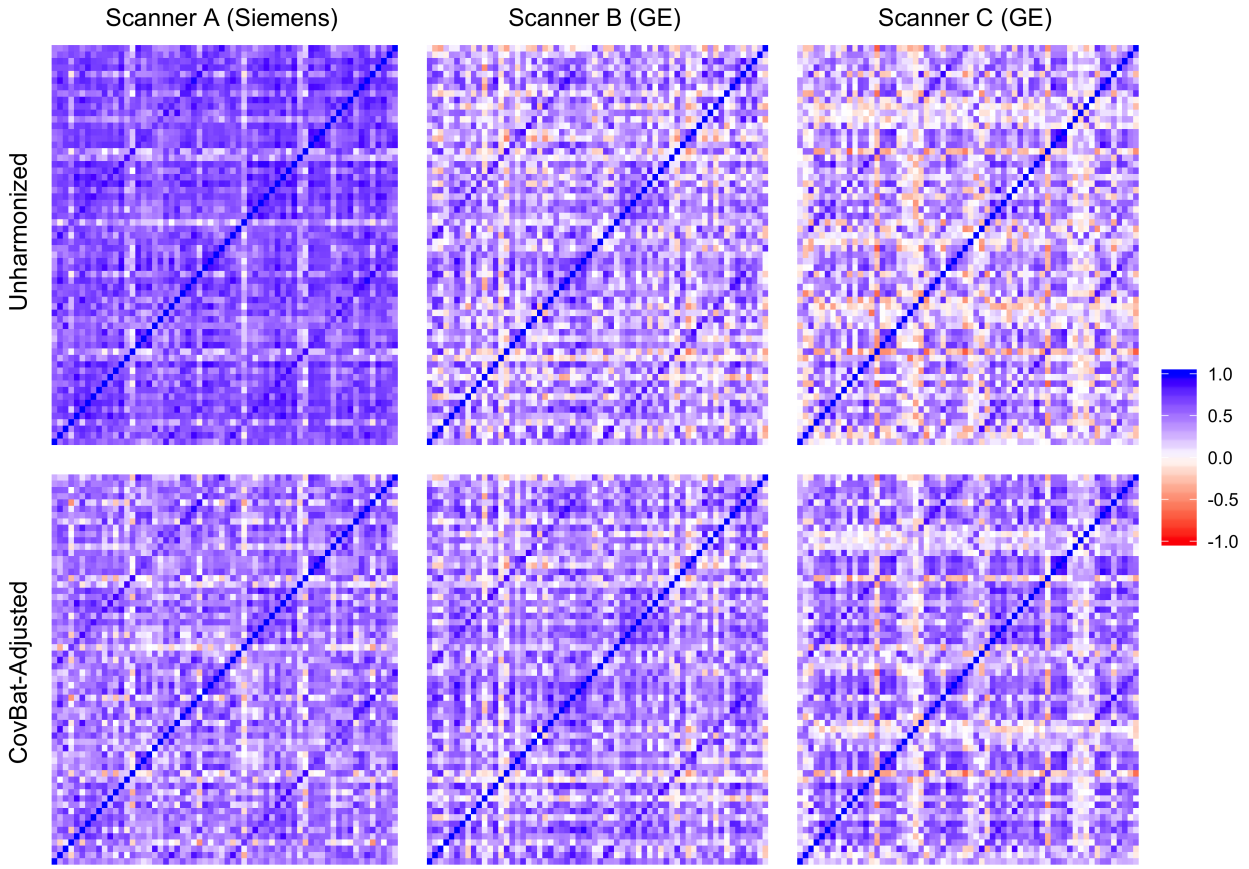

Figure 2: Correlation matrices for measurements acquired on each scanner before and after CovBat harmonization. ComBat-adjusted matrices are visually indistinguishable from unharmonized

	Original	ComBat	CovBat
A,B	507.3	507.3	221.0 (218.4 - 224.7)
A,C	760.9	760.9	268.3 (267.3 - 277.6)
B,C	308.6	308.6	258.2 (256.6 - 259.8)

Table 1: Pairwise Frobenius norms between scanner-specific correlation matrices. Results from adjusting the number of PCs employed ranging from those required to explain 50% to 99% of variation are reported in parentheses as the minimum and maximum pairwise Frobenius norms across the range.

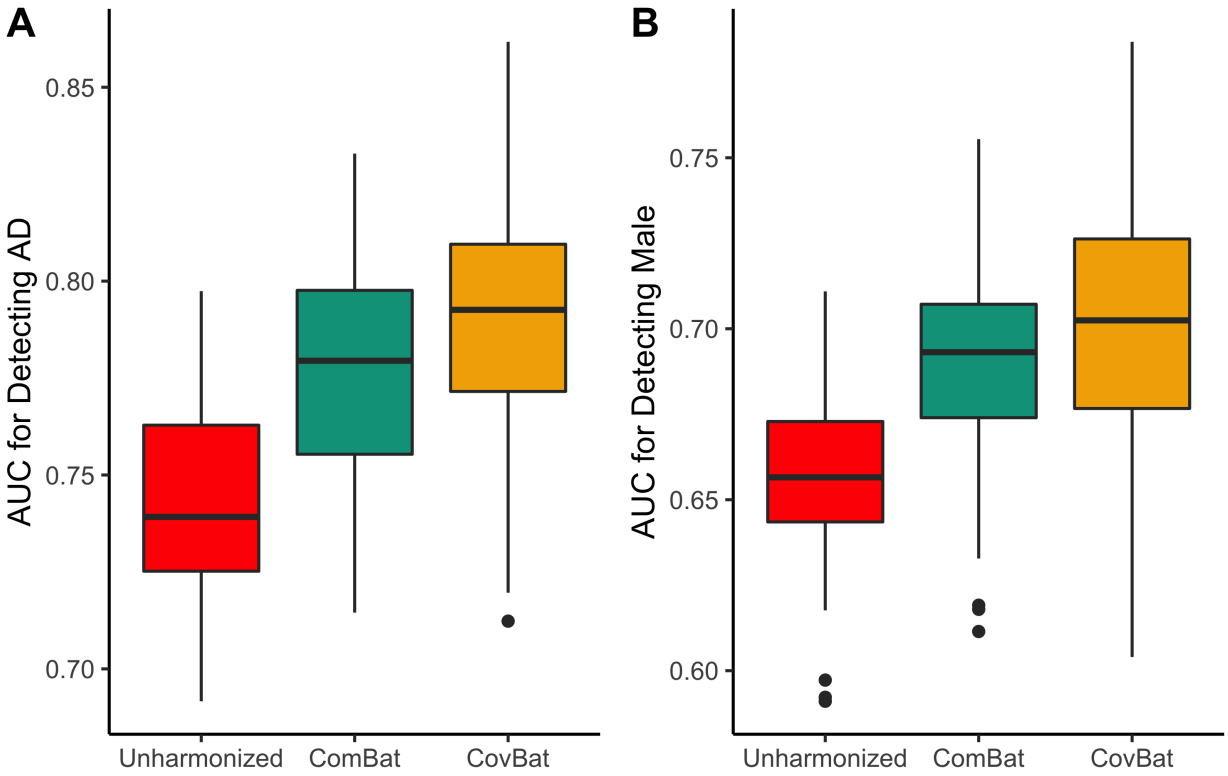

Figure 3: Results from MVPA experiments for detection of Alzheimer's disease status and male sex using cortical thickness data. The data is randomly split into 50% training and 50% validation then used to train a random forests algorithm to predict either trait. AUC values over 100 repetitions of this experiment are reported for unharmonized, ComBat-adjusted, and CovBat-adjusted data. Results for detection of Alzheimer's disease status are shown in (A) and results for detection of male are shown in (B).

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. Figure 3 shows 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 0.74 ($\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$0.04). These findings suggest that CovBat not only provides thorough removal of scanner effects, but also helps to recover clinical associations.

Findings Replicated in Simulations

To test our harmonization method, we create simulated datasets based on a modified version of the ComBat model which includes scanner effects in covariance. We impose mean and variance scanner effects on a ground truth multivariate normal distribution and additionally modify the covariance matrix of this distribution by scanner. To achieve the latter, we add high-rank scanner-specific matrices to the underlying true covariance matrix to ensure that the scanner effect can be corrected through adjustment in PC scores, but also requires harmonization of a sufficiently high number of PCs. To test detection of a simulated covariate, we impose that distribution of the outcome measures depends on the presence of a binary covariate drawn from a Bernoulli distribution. Our simulations consisted of three scanners each with 100 simulated subjects with a binary covariate drawn from a Bernoulli(0.25) distribution. We further simulate a covariate associated with a decrease in the mean of the cortical thickness values for 15 ROIs in both hemispheres. Additional details are available in *SI Appendix*.

Covariate Effect on Mean

In a first scenario where the covariate does not influence the covariance, we anticipate that harmonization of mean and variance is sufficient to remove any confounding of scanner effect with the association of interest. We also anticipate that detection of scanner via MVPA should be improved by harmonization of covariance. We perform experiments to test these hypotheses under the same paradigm as the MVPA experiments implemented on the ADNI dataset and report the results in Figure 4. The results show that both ComBat (AUC 0.59 ± 0.04) and CovBat (0.54 ± 0.04) underperform compared with the raw data (0.63 ± 0.05) for detection of the simulated covariate. This result could be attributable to covariate effects remaining after the ComBat residualization step, which would then be removed through the harmonization of mean across scanners. As for detection of scanner, we find that scanner 1 is almost perfectly detected in the raw data (AUC 0.999 ± 0.001), obscured after ComBat (0.58 ± 0.05) and nearly impossible to detect after CovBat (0.53 ± 0.03).

Covariate Effect on Covariance

In a second simulation scenario, we study the impact of an additional covariate effect on variance and covariance that is confounded with the scanner effects. To achieve this, we allowed the covariate effect on covariance to be proportional to a chosen scanner's covariance shift (see *SI Appendix* for details). This scenario represents a situation where detection of the covariate using MVPA would be highly influenced by the presence of scanner effects. Without harmonization of covariance, observations from the chosen scanner resemble observations obtained from subjects with the covariate. Consequently, we expect that ComBat alone would be insufficient to recover the covariate association and that CovBat would outperform on this metric. Since we made no changes to the scanner effects, we again anticipate that detection of scanner should become less accurate after CovBat. The results of the MVPA experiments are shown in Figure 4. As anticipated, we observe that the mean AUC using the raw data is the lowest (0.82 ± 0.03), ComBat shows some performance increase, (0.85 ± 0.03),

and CovBat performs the best (0.88±0.02). Detection of scanner also follows our observations in ADNI data with Scanner 1 almost perfectly detected in the raw data (AUC 0.999±0.001), difficult to detect after ComBat (0.58 ± 0.04) and nearly impossible to detect after CovBat (0.53 ± 0.03).

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 further replicated these observations, and suggest that the improvement in covariate detection could be linked to confounding between scanner effect and covariate effect on the covariance between multivariate measurements. This finding suggests that future work could aim to control for covariate effects on variance and covariance so that harmonization does not remove desired properties of the data. While our study focused on structural neuroimaging data, our findings extend 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.

Acknowledgements

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

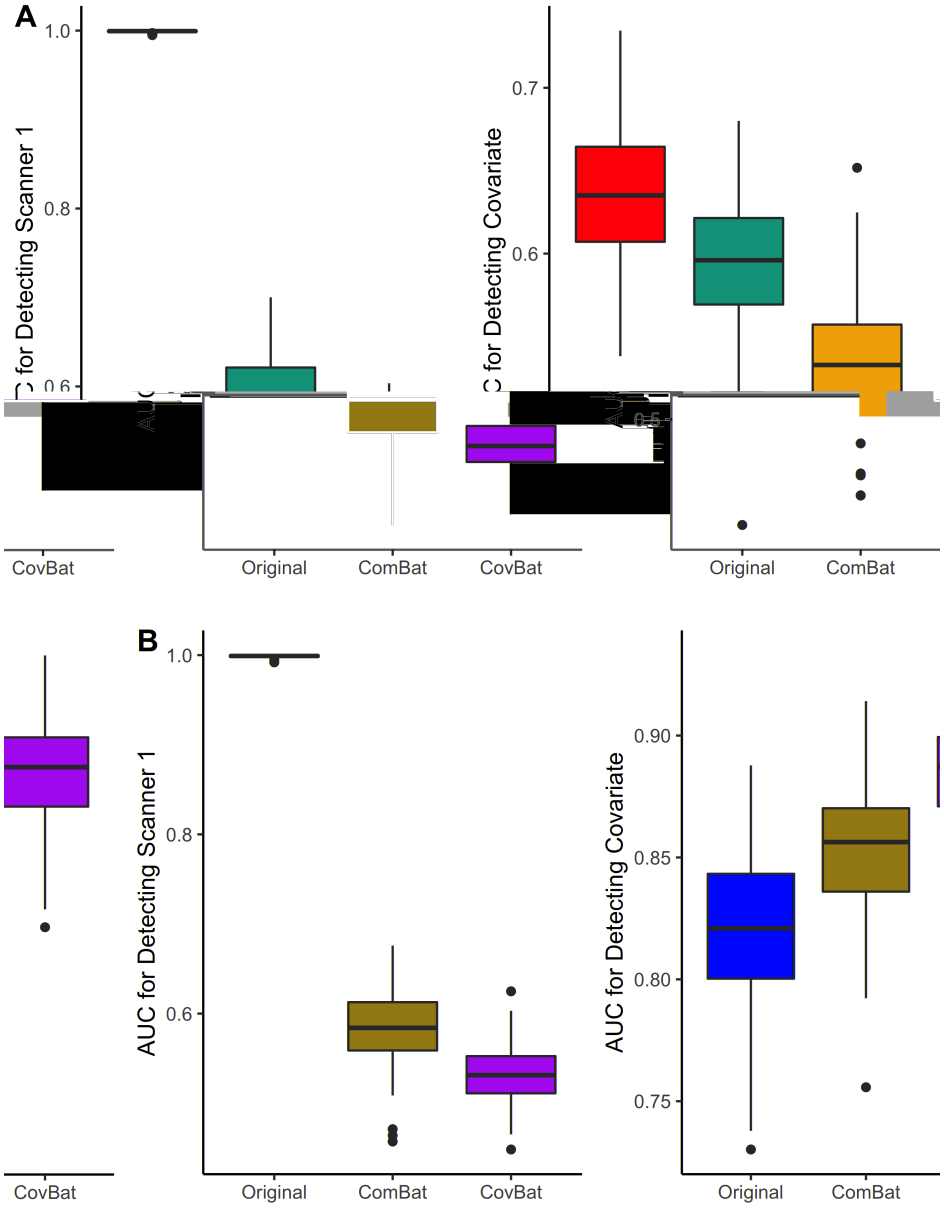

Figure 4: Simulation MVPA experiment results for detection of scanner and for detection of a binary covariate. The data is randomly split into 50% training and 50% validation then used to train a random forests algorithm to predict either Scanner 1 or the presence of the binary covariate. AUC values over 100 repetitions of this experiment are reported for unharmonized, ComBat-adjusted, and CovBat-adjusted data. AUC values are shown for detection of scanner and covariate under covariate effect on mean only (A) and mean, variance, and covariance (B)

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.

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.

Jack, C. R., Bernstein, M. A., Borowski, B. J., Gunter, J. L., Fox, N. C., Thompson, P. M., Schuff, N., Krueger, G., Killiany, R. J., DeCarli, C. S., Dale, A. M. & Weiner, M. W. (2010). Update on the MRI Core of the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & dementia : the journal of the Alzheimer's Association* **6**, 212–220.

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

Klein, A. & Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience* **6**, 171.

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., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., Kandel, B. M., van Strien, N., Stone, J. R., Gee, J. C. & Avants, B. B. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage* **99**, 166–179.

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. (2018). The ANTs Longitudinal Cortical Thickness Pipeline. *bioRxiv* , 170209.

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.

Supporting Information Appendix (SI)

ADNI Dataset Demographics

The subsample of 505 subjects included in the study have 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. For the subsample comprised of the three largest sites in the dataset, their demographics are listed in Table 3. Since the correlations between cortical thickness values are of primary interest in our study, we display the correlation matrices annotated with the 62 regions of interest (ROIs) in Figure 6.

Harmonization using 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 both analyses 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 β_v, $v = 1, 2, \ldots, 62$ using only the training subjects. We report the results in Figure 7. The results appear quite similar to harmonization using the full dataset, except with additional variance in the AUC values for detection of male. Detection of site still worsens after ComBat (AUC 0.67 ± 0.03) and is almost at chance after CovBat (AUC 0.54 ± 0.03). For detection of AD, improvements are demonstrated after ComBat adjustment (AUC 0.76 ± 0.03) and greater improvements after CovBat (AUC 0.77 ± 0.03). For detection of male, lesser improvement is observed from ComBat (AUC 0.68 ± 0.03) to CovBat (AUC 0.68 ± 0.03).

Simulation Settings

In the first simulation setting, we assume that the covariate only affects the mean of the measurements. We choose $\beta_v = -0.5$ 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 covariate. We also choose $\Psi = \mathbf{0}$ where $\mathbf{0}$ is a 62×62 zero matrix to ensure that each site-specific covariance matrix only depends on the underlying true covariance matrix Σ and the chosen Ω_i matrices. The covariance matrices across sites are shown in Figure 5 with associated pairwise Frobenius norms listed in Table 2.

In the second simulation setting, we assume that the covariate affects not only mean, but also variance and covariance. To achieve this, we use the same β value as above but

	Original	ComBat	CovBat
A,B	365.73	369.88	153.58
A,C	185.33	183.38	115.95
B,C	168.79	173.39	127.10

Table 2: Pairwise Frobenius norms between site-specific correlation matrices for simulations with no effect of covariate on covariance.

	A (Siemens)	B (GE)	C (GE)	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 3: ADNI demographics by scanner for the three scanners with the largest number of acquired subjects. ANOVA p-values for testing for a difference in mean across groups are reported in the rightmost column.

choose Ψ to be related to Ω_2 to force confounding of site and covariate effects on covariance. To achieve this, we have $\Psi_{i,i} = \Omega_{i,i}$ and $\Psi_{i,j} = -\Omega_{i,j}/2$ for $i \neq j$ and $i = 1, 2, \ldots, 62$, $j = 1, 2, \ldots, 62$. The simulation findings are shown in the main paper be consistent with findings from the ADNI data application. To better illustrate the effects of harmonization, we plot the stratified covariance matrices for subjects whose binary covariate equals 0 or 1, before and after CovBat in Figure 8. We observe that CovBat harmonization leads to better differentiation between the two subject groups. Meanwhile the differences across sites are much smaller after CovBat as evident in Figure 9 and Table 4.

	Original	ComBat	CovBat
A,B	297.25	297.79	126.79
A,C	122.26	121.31	127.91
B,C	263.50	266.10	128.67

Table 4: Pairwise Frobenius norms between site-specific correlation matrices for simulations with effect of covariate on covariance.

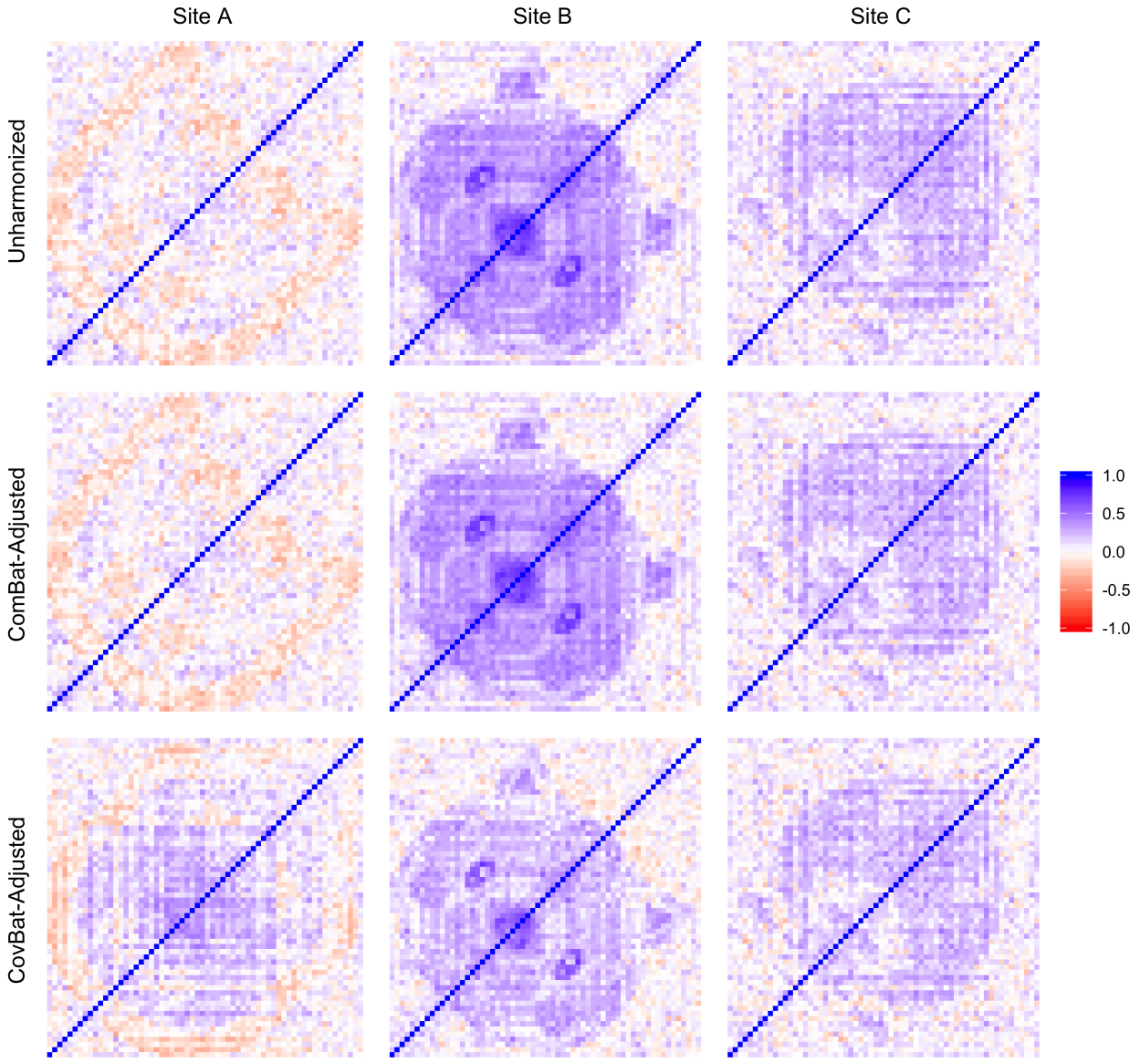

Figure 5: Correlation matrices following harmonization for simulations with no effect of covariate on covariance.

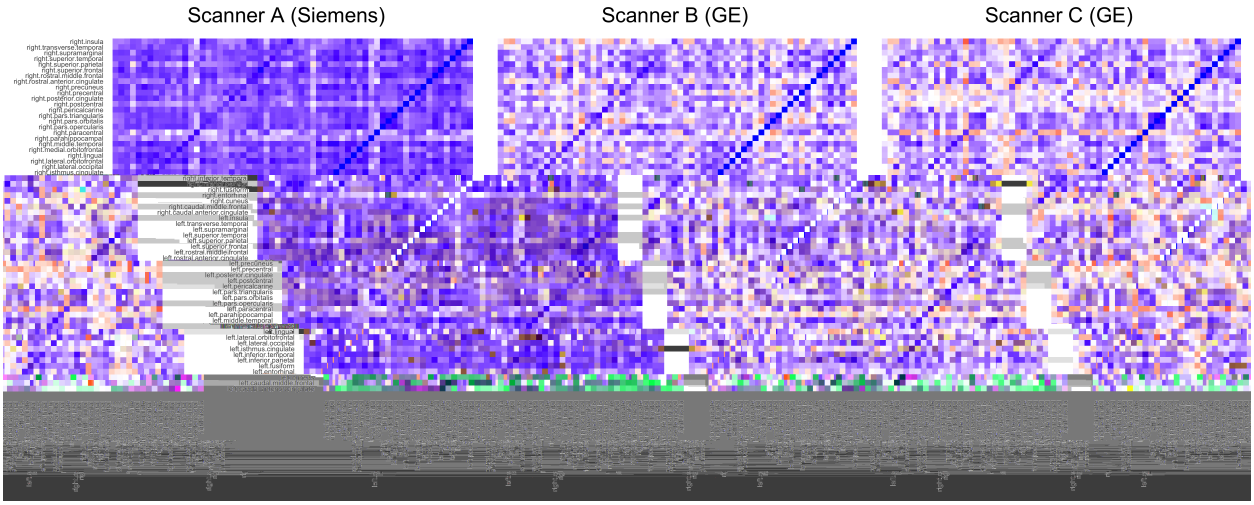

Figure 6: Correlation matrices for each scanner. **(A)** shows a scanner manufactured by Siemens while (B) and (C) were manufactured by General Electric.

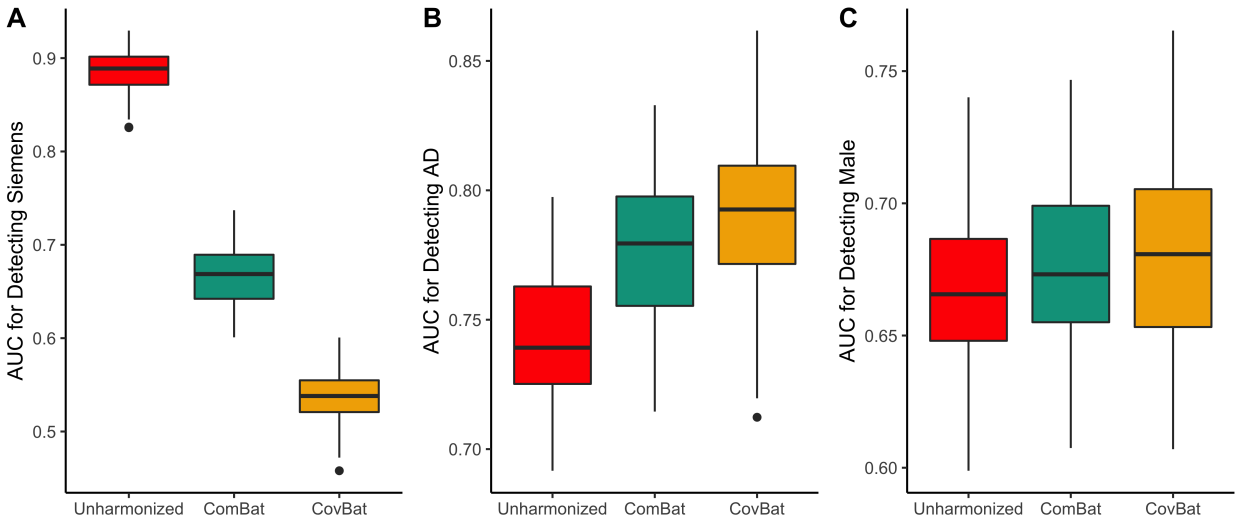

Figure 7: MVPA experiment results for harmonization using only training data. The data is randomly split into 270 training subjects and 235 testing subjects such that every site is represented in each group. The training set is then used to train a random forests algorithm to predict scanner or patient 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).

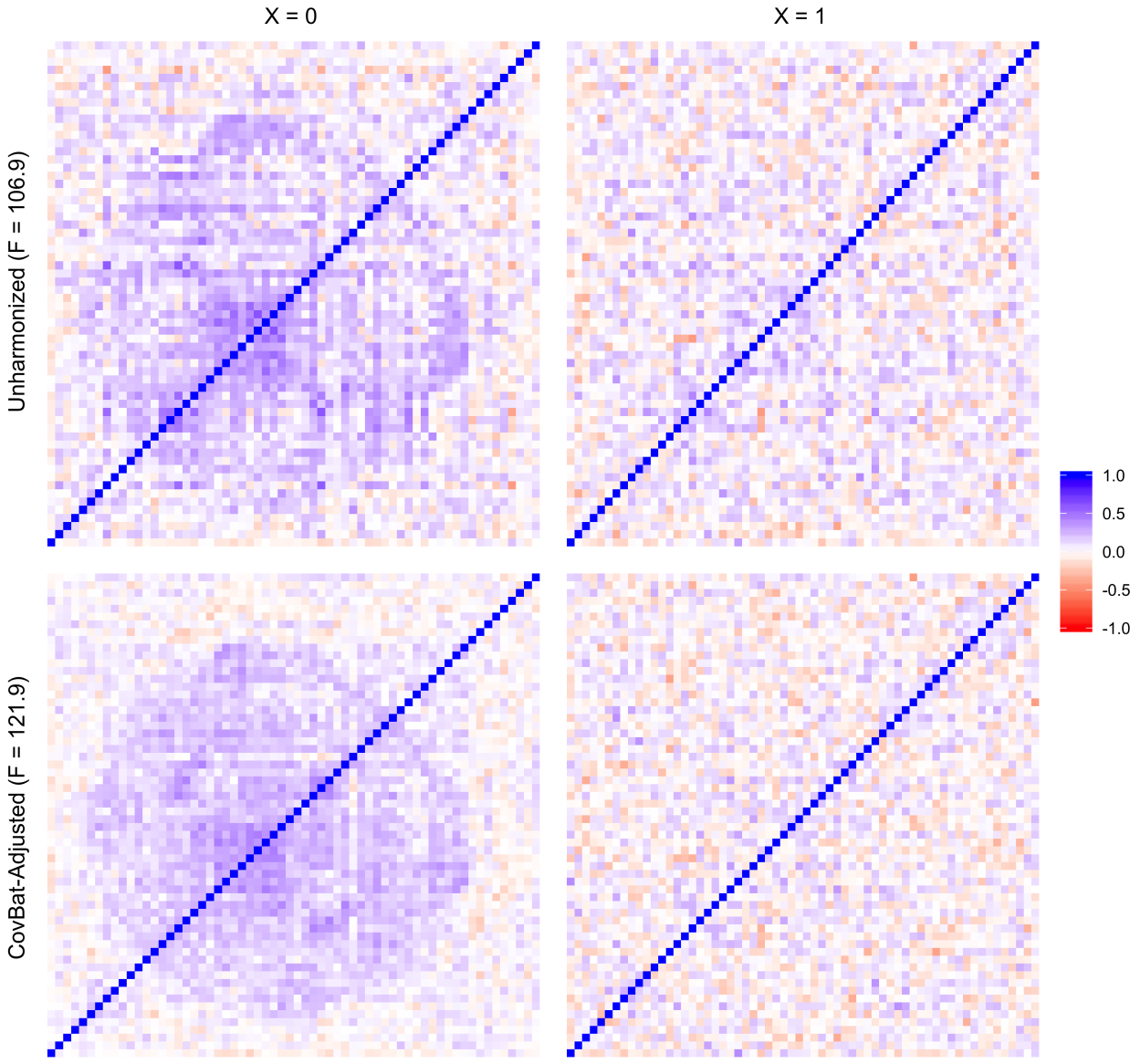

Figure 8: Correlation matrices across levels of covariate before and after CovBat harmonization in simulations with effect of covariate on covariance.

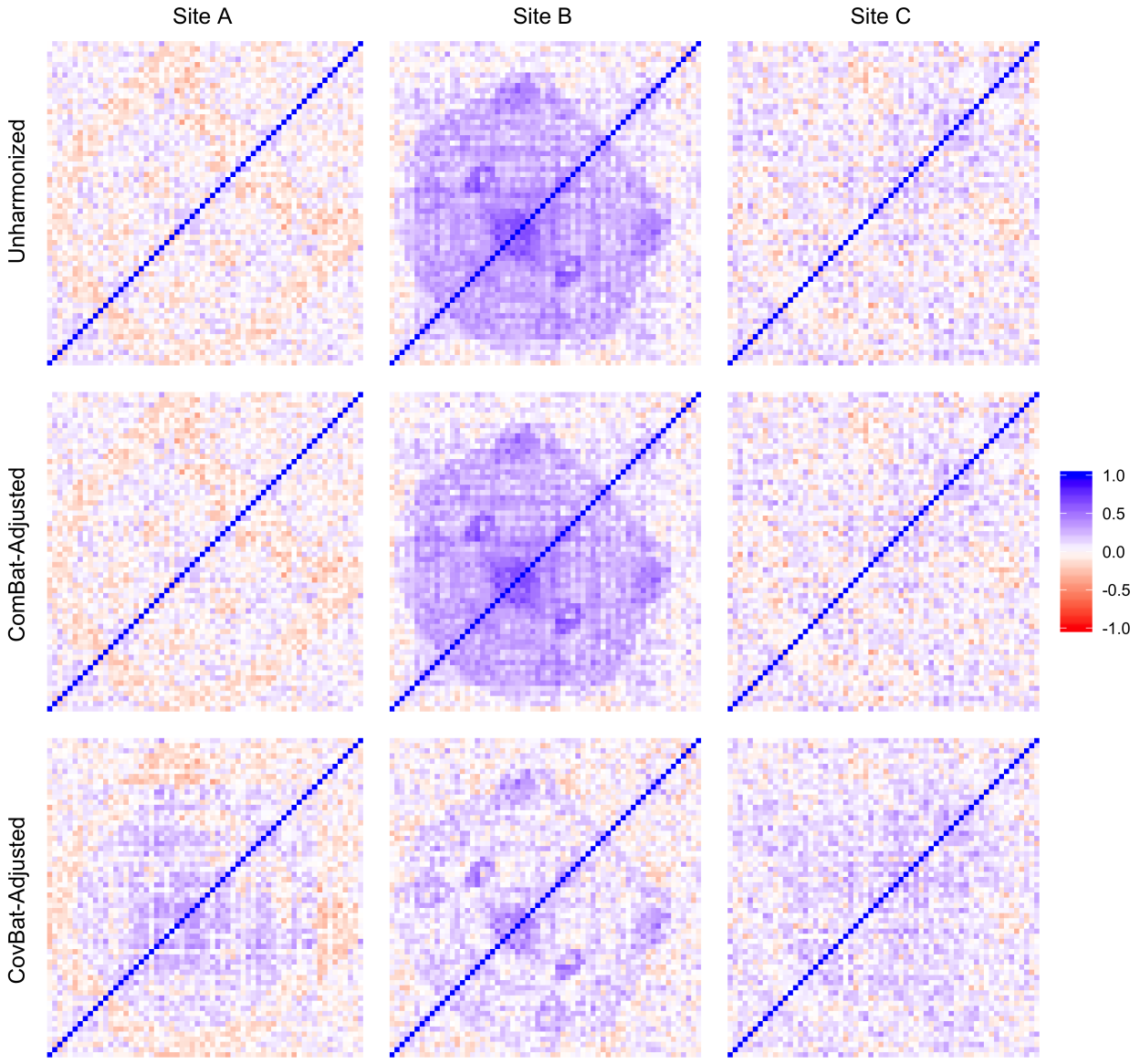

Figure 9: Correlation matrices following harmonization for simulations with effect of covariate on covariance.