Imaging Neuroscience
an open access journal

Research Article

RELIEF: A structured multivariate approach for removal of latent inter-scanner effects

Rongqian Zhang[a], **Lindsay D. Oliver**[b], **Aristotle N. Voineskos**[b,c], **Jun Young Park**[a,d]

[a]Department of Statistical Sciences, University of Toronto, Toronto, Canada
[b]Centre for Addiction and Mental Health, Toronto, Canada
[c]Department of Psychiatry, University of Toronto, Toronto, Canada
[d]Department of Psychology, University of Toronto, Toronto, Canada

Corresponding Author: Jun Young Park (junjy.park@utoronto.ca)

ABSTRACT

Combining data collected from multiple study sites is becoming common and is advantageous to researchers to increase the generalizability and replicability of scientific discoveries. However, at the same time, unwanted *inter-scanner biases* are commonly observed across neuroimaging data collected from multiple study sites or scanners, rendering difficulties in integrating such data to obtain reliable findings. While several methods for handling such unwanted variations have been proposed, most of them use univariate approaches that could be too simple to capture all sources of scanner-specific variations. To address these challenges, we propose a novel multivariate harmonization method called RELIEF (**RE**moval of **L**atent **I**nter-scanner **E**ffects through **F**actorization) for estimating and removing both explicit and latent scanner effects. Our method is the first approach to introduce the simultaneous dimension reduction and factorization of interlinked matrices to a data harmonization context, which provides a new direction in methodological research for correcting inter-scanner biases. Analyzing diffusion tensor imaging (DTI) data from the Social Processes Initiative in Neurobiology of the Schizophrenia (SPINS) study and conducting extensive simulation studies, we show that RELIEF outperforms existing harmonization methods in mitigating inter-scanner biases and retaining biological associations of interest to increase statistical power. RELIEF is publicly available as an R package.

Keywords: batch effects, covariance heterogeneity, dimension reduction, inter-scanner biases, neuroimaging, RELIEF

1. INTRODUCTION

It is increasingly common in neuroimaging and genomics to combine data collected from multiple study sites to increase the power and the reproducibility of scientific discoveries. However, combining such data comes with unwanted non-biological variations that need to be removed for successful data integration. In neuroimaging, this is often characterized by inter-scanner biases (scanner effects) when subject data are obtained by using different magnetic resonance imaging (MRI) scanners with different optimization protocols. These inter-scanner biases have been shown to be present in most neuroimaging data types, including diffusion (Vollmar et al., 2010; Zhu et al., 2011), structural (Han et al., 2006; Takao, Hayashi, & Ohtomo, 2014), and functional (Dansereau et al., 2017) MRI. These terms are analogous to *batch effects* in genomic studies that are observed with genome-wide microarray or RNA sequencing data with different sample preparation and sequencing methods.

There have been numerous efforts in statistics, such as ComBat, to capture and remove these unwanted variations and increase the signal-to-noise ratio (J.-P. Fortin

Received: 27 June 2023 Accepted: 2 August 2023 Available Online: 11 August 2023

Imaging Neuroscience, Volume 1, 2023
https://doi.org/10.1162/imag_a_00011

et al., 2017, 2018; Johnson, Li, & Rabinovic, 2007; M. Yu et al., 2018; Y. Zhang, Parmigiani, & Johnson, 2020). ComBat (Johnson et al., 2007) is a popular regression-based batch correction approach first motivated from microarray data, and has been promising in removing inter-scanner biases in many neuroimaging data types, including fractional anisotropy and mean diffusivity (J.-P. Fortin et al., 2017), cortical thickness (J.-P. Fortin et al., 2018), and functional connectivity (M. Yu et al., 2018). In ComBat, scanner effects are characterized by an additive scanner effect (location) and a multiplicative scanner effect (scale) for each feature. While a regression model is used in each feature, ComBat uses empirical Bayes to stabilize estimates across features and provides robustness in the case of small within-scanner sample sizes (Johnson et al., 2007). In addition to showing its utility in various neuroimaging data types, ComBat has been extended to harmonize imaging data collected in a longitudinal manner (Beer et al., 2020), to preserve non-linear age trajectories of cortical thickness data in mega-analysis in cross-sectional studies (Pomponio et al., 2020). It is also a versatile method that allows for harmonization even without the need to share original data from a study site with other sites, which relaxes concerns about data privacy (Chen, Luo, et al., 2022).

The ComBat's location-scale model is simple and interpretable, but, from the statistical perspective, it is insufficient to capture all sources of scanner effects. The heterogeneity in *covariances* across different sites or scanners has been overlooked in the neuroimaging literature, and such heterogeneity might also lead to decreased statistical power. ComBat is oversimplified by the assumption that additive scanner effects can be explained by only an intercept for each scanner and feature. Recently, a new harmonization method called CovBat (Chen, Beer, et al., 2022) was proposed to address covariance heterogeneity in multi-site, multi-scanner studies by extending ComBat. It applies ComBat twice: first to the original data, then to the principal component scores from the residual matrix. CovBat is an important development that expanded the scope of statistical harmonization to address heterogeneous covariances, and it has been shown to be more efficient than ComBat, as expected (Chen, Beer, et al., 2022; Chen, Srinivasan, et al., 2022). However, CovBat implicitly assumes that the covariance scanner effect is contained within the eigenspace of the residual matrix only, in the form of a location-scale model. As Chen, Beer, et al. (2022) noted, this assumption may limit the ability of CovBat to characterize all sources of covariance heterogeneity, which we also show in this paper.

The method for harmonizing covariances across scanners can be understood using the latent variable formulation (Chen, Beer, et al., 2022). Singular value decomposition (SVD) and principal component analysis (PCA) are commonly used techniques for removing or adjusting for non-biological variations not explicitly specified by scanner information. SVA (Surrogate Variable Analysis) is a method that was originally developed for genomic studies (Leek & Storey, 2007) and then adapted to neuroimaging studies (J.-P. Fortin et al., 2016). SVA includes latent factors of unwanted variation as surrogate variables, which are not associated with the biological covariates of interest. Instead of using explicit variables to denote scanner effects, SVA identifies and estimates scanner or other non-biological artifacts through permutation testing, then removes them as surrogate variables. RAVEL (J.-P. Fortin et al., 2016) is a statistical method for correcting technical variability in neuroimaging data. RAVEL applies SVD to obtain latent factors of unwanted variations in the control regions and then removes the latent factors and corresponding effects in the test regions (J.-P. Fortin et al., 2016, 2017). These approaches that apply low-rank factorization methods to all study subjects' imaging features are fundamentally limited to addressing scanner-specific latent effects. At the same time, efforts to identify low-rank factors for study subjects from the same scanner may overkill biological variations.

In this paper, we propose a novel harmonization method called RELIEF (**RE**moval of **L**atent **I**nter-scanner **E**ffects through **F**actorization) to distinguish loadings shared across scanners (which should be preserved) from loadings specific to scanners (which should be removed), which enhances the current understanding of inter-scanner biases. We formulate latent scanner effects from the perspective of linked matrix factorization by extending the recent work of Park and Lock (2020) in the harmonization context. It aligns with growing methodological developments on simultaneous dimension reduction and factorization of multi-modal data (e.g., Gaynanova & Li, 2019; Lock, Hoadley, Marron, & Nobel, 2013; Lock, Park, & Hoadley, 2022), which has also been shown to be promising in neuroimaging data (Q. Yu, Risk, Zhang, & Marron, 2017). Through extensive data analyses and simulations, we show our proposed method has superior performance in identifying and removing latent unwanted variations specific to each scanner, thus leading to covariance homogeneity across scanners and increasing statistical power compared to existing methods. Also, our estimation procedure is scalable and takes only a few seconds to implement, which supports its practical utility.

The rest of the paper is organized as follows. Section 2 describes our proposed method, RELIEF, and compares it to existing harmonization methods. In Section 3, we apply our method to the fractional anisotropy (FA) and mean diffusivity (MD) data from the Social Processes Initiative in the Neurobiology of the Schizophrenia(s) (SPINS) study, where study subjects were collected from multiple sites and scanners. We compare RELIEF to other harmonization methods using a comprehensive evaluation framework. Section 4 conducts extensive simulations to evaluate performances in terms of Type 1 error rate and statistical power. We conclude with some points of discussion in Section 5.

2. METHODS

2.1. Notation and setup

We let $i=1,\ldots,M$ denote the index for each scanner (batch), $j=1,\ldots n_i$ denote the subject index in ith scanner ($\sum_{i=1}^{M} n_i = n$), and $v=1,\ldots,V$ denote the index for imaging features. We let $\mathbf{x}_{ij}$ be the q-dimensional covariate vector for jth subject in ith scanner (e.g., age and sex). y_{ijv} is the vth imaging feature of the jth subject of the ith scanner. By stacking all observations of $\mathbf{x}_{ij}$, we let $\mathbf{X}$ be a $n \times q$ matrix of q covariates observed for n study subjects. Similarly, let $\mathbf{Y}$ be a $V \times n$ data matrix of V features. Then, to group the subjects from the same scanner together, we consider $\{\mathbf{Y}_i : V \times n_i \mid i=1,\ldots,M\}$ a partition of $\mathbf{Y}$. The matrices can be concatenated to form a matrix $\mathbf{Y}=[\mathbf{Y}_1;\mathbf{Y}_2;\ldots;\mathbf{Y}_M]$. We will use this notation for a general $V \times n$ matrix throughout this paper.

2.2. Existing harmonization methods

2.2.1. Adjusted residuals (AdjRes)

The simplest approach to model inter-scanner bias is to use a regression-based approach to characterize additive scanner-specific deviations for each feature. AdjRes considers the following specifications,

$$y_{ijv} = \alpha_v + \mathbf{x}_{ij}' \boldsymbol{\beta}_v + \gamma_{iv} + \epsilon_{ijv}, \tag{1}$$

where, for the vth feature, α_v is the intercept, $\boldsymbol{\beta}_v$ is the regression coefficients for $\mathbf{x}_{ij}$, and ϵ_{ijv} is a Gaussian noise. The parameters α_v, $\boldsymbol{\beta}_v, \gamma_{iv}$ can be estimated by the least squares method. The scanner-specific means, γ_{iv}, needs to be removed and the harmonized data are constructed by $y_{ijv}^{AdjRes} = \hat{\alpha}_v + \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}_v + \hat{\epsilon}_{ijv}$.

2.2.2. ComBat

ComBat seeks to remove the additive and multiplicative scanner effects (Johnson et al., 2007). For the vth feature, ComBat characterizes the additive and multiplicative scanner effects by

$$y_{ijv} = \alpha_v + \mathbf{x}_{ij}' \boldsymbol{\beta}_v + \gamma_{iv} + \phi_{iv}\epsilon_{ijv}. \tag{2}$$

In Equation (2), the scanner effects are characterized by γ_{iv} (the additive scanner effect) and ϕ_{iv} (the multiplicative scanner effect). After obtaining $\hat{\alpha}_v$, $\hat{\boldsymbol{\beta}}_v$ via least squares, ComBat estimates scanner effects in locations (i.e., $\gamma_{iv}^{\star}$) and scales (i.e., $\phi_{iv}^{\star}$) via empirical Bayes for each feature separately, providing stable and robust estimations of these parameters in the case of small within-scanner sample sizes (Johnson et al., 2007). The ComBat-harmonized data is defined by $y_{ijv}^{ComBat} = \hat{\alpha}_v + \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}_v + \hat{\epsilon}_{ijv}^{ComBat}$, where

$$\hat{\epsilon}_{ijv}^{ComBat} = \frac{y_{ijv} - \hat{\alpha}_v - \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}_v - \gamma_{iv}^{\star}}{\phi_{iv}^{\star}}. \tag{3}$$

2.2.3. CovBat

In addition to ComBat's model in Equation (2), CovBat assumes that the error terms $\epsilon_{ij} = (\epsilon_{ij1}, \epsilon_{ij2}, \ldots, \epsilon_{ijV})' \sim \mathcal{MVN}(\mathbf{0}, \Sigma_i)$, where Σ_i is the covariance for the ith scanner. CovBat further assumes the underlying pooled covariance is homogeneous across scanners. Inspired by how ComBat mitigates the difference between the variance within each scanner and the pooled variance, CovBat shifts the within-scanner covariance to the pooled covariance by using principal component (PC) and PC scores. CovBat's harmonization procedure is summarized as follows. First, ComBat is applied to full imaging data, yielding ComBat-residuals as in Equation (3) with homogeneous variances across scanners. CovBat then conducts the eigendecomposition on the sample covariance of Combat-residuals and applies ComBat again to principal component scores to remove heterogeneous means and variances, which yields CovBat-residuals with an additional source of scanner effect removed. The final CovBat-harmonized data is $y_{ijv}^{CovBat} = \hat{\alpha}_v + \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}_v + \hat{\epsilon}_{ijv}^{CovBat}$. CovBat assumes that the covariance scanner effects can be captured by the location-scale adjustments to the principal components of the residuals. Despite its efficiency, we point out that CovBat's assumption might not be sufficient to characterize all sources of covariance heterogeneity.

2.3. New method: RELIEF (REemoval of Latent Inter-scanner Effects through Factorization)

We first characterize three sources of scanner effects (additive mean (location), additive latent, and multiplicative scanner effects (scale)) via an additive multivariate model illustrated in Figure 1. We assume that the data matrix **Y** consists of

$$\mathbf{Y} = \mathbf{A} + \beta\mathbf{X}' + [\Gamma_1;\ldots;\Gamma_M] + \mathbf{R} + [\mathbf{I}_1;\ldots;\mathbf{I}_M] + [\delta_1\mathbf{E}_1;\ldots;\delta_M\mathbf{E}_M], \tag{4}$$

where **A** is the intercept matrix (rank of 1), β is a $V \times q$ matrix of regression coefficients (rank of $\min(V, q)$), and $[\Gamma_1;\ldots;\Gamma_M]$ is a matrix of additive scanner effects (locations) for each feature (rank of M), where elements of each row of Γ_i take the same value. Note that $\mathbf{A} + \beta\mathbf{X}' + [\Gamma_1;\ldots;\Gamma_M]$ in Equation (4) corresponds to the collection of $\alpha_v + \mathbf{x}'_{ij}\beta_v + \gamma_{iv}$ in Equation (1) across all imaging features.

The RELIEF model assumes that ϵ_{ijv} in Equation (1) is decomposed into three additive variations. Specifically:

- **R** is a $V \times n$ matrix of the latent structure explaining shared variations across all scanners but not explained by covariate effects. It includes (i) non-linear covariate effects from **X** or (ii) any additional variations due to unobserved covariates. From the viewpoint of scanner-effect correction, this should be preserved after harmonization.
- $\mathbf{I}_i$ is a $V \times n_i$ matrix of latent variations explaining latent scanner effects in the ith scanner beyond scanner-specific means Γ_i (locations). This might include any non-linear scanner effects (Cetin-Karayumak et al., 2020). It should be removed after harmonization.
- $\delta_i\mathbf{E}_i$ is a $V \times n_i$ noise matrix, and each element of $\mathbf{E}_i$ is assumed have a unit variance. δ_i characterizes the variance heterogeneity as specified in ComBat, which has shown to be promising in neuroimaging. From the viewpoint of scanner-effect correction, δ_is should be standardized to have a common variance across scanners.

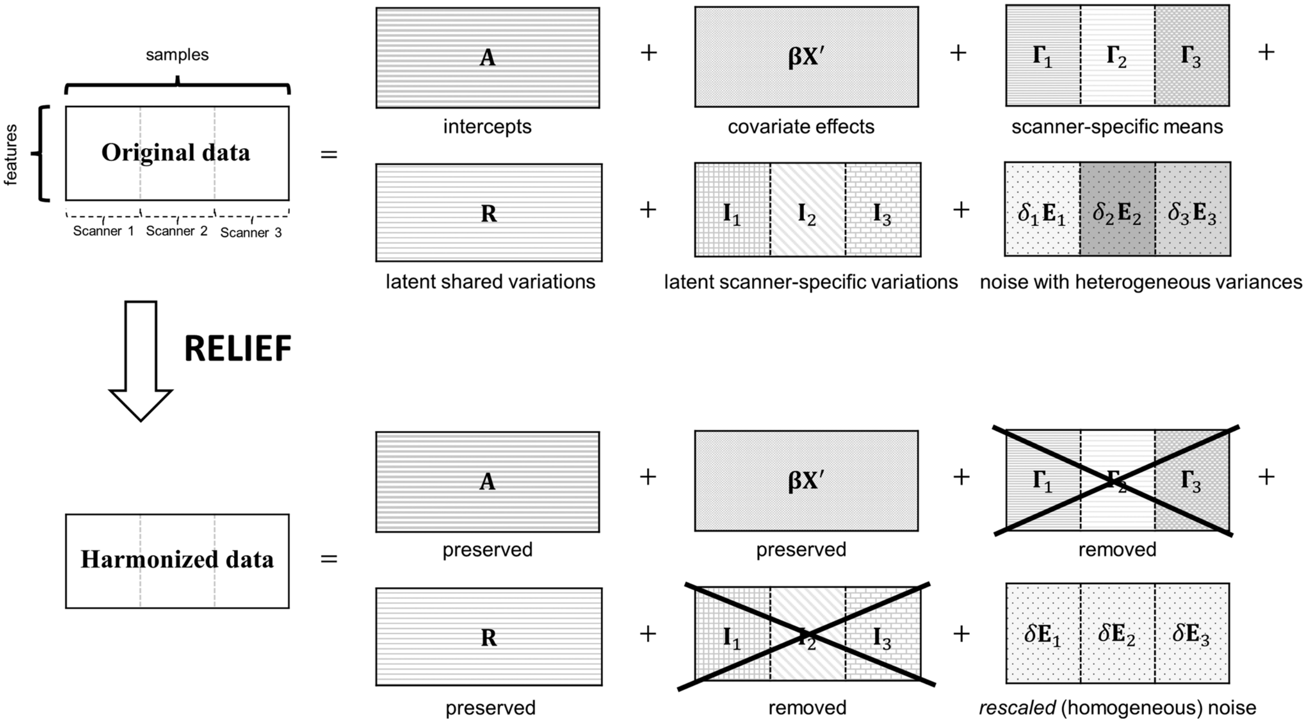

Fig. 1. Overview of RELIEF using data consisting of three scanners for illustrations. It decomposes original data as (i) covariate effects, (ii) scanner-specific means (locations), (iii) latent shared variations, (iv) latent scanner-specific variations, and (v) noise with heterogeneous variances (scales). For harmonization purposes, RELIEF removes (ii) and (iv) specific to scanners and homogenizes (v).

Throughout this paper, we assume $\mathbf{R}$ as well as each of $\mathbf{I}_1,\ldots,\mathbf{I}_M$ to be low rank, and estimate their ranks using a model-based approach.

Our approach is summarized by (i) removing scanner, feature-specific means and obtaining covariate effects first, (ii) standardizing the data matrix to have homogeneous variance, (iii) decomposing it into scanner-specific and scanner-independent factors, and (iv) reconstructing harmonized data.

Steps (i) and (ii) are achieved through the preprocessing step. We obtain $\hat{\mathbf{A}}$, $\hat{\beta}$ and $[\hat{\Gamma}_1;\ldots;\hat{\Gamma}_M]$ by using the two-step regression. Specifically, we first fit GLM using the intercept and covariates (A and β) and obtain residuals $(\mathbf{Y}-\hat{\mathbf{A}}-\hat{\beta}\mathbf{X}')$. Then, using the residuals from the first step, we remove scanner-specific means for each feature (Γ) to obtain the second-step residuals $[\mathbf{Y}_1-\hat{\mathbf{A}}_1-\hat{\beta}\mathbf{X}_1'-\hat{\Gamma}_1;\ldots;\mathbf{Y}_M-\hat{\mathbf{A}}_M-\hat{\beta}\mathbf{X}_M'-\hat{\Gamma}_M]$ to be used in subsequent steps. When the variability of the second-step residuals differs across features, we can easily scale each residual by its residual standard deviation, apply steps (iii) and (iv), and scale back each feature.

Step (iii) is achieved by simultaneous dimension reduction and factorization methods proposed by Park and Lock (2020) and Lock et al. (2022). We first scale each residual matrix from the last step by $\hat{\delta}_i$ in order to make the residual variances homogeneous across $i=1,\ldots,M$: $\Delta\equiv[(\mathbf{Y}_1-\hat{\mathbf{A}}_1-\hat{\beta}\mathbf{X}_1'-\hat{\Gamma}_1)/\hat{\delta}_1;\ldots;(\mathbf{Y}_M-\hat{\mathbf{A}}_M-\hat{\beta}\mathbf{X}_M'-\hat{\Gamma}_M)/\hat{\delta}_M]$. Following Park and Lock (2020) and Lock et al. (2022), we estimate $\hat{\delta}_i$ by the median of the singular values of residual matrices for each scanner divided by the square root of the median of the Marcenko–Pastur distribution (Gavish & Donoho, 2017). Provided that $\hat{\delta}_i\approx\delta_i$, we first note that Δ is represented by

$$\Delta=\mathbf{R}^{\star}+\mathbf{I}^{\star}+\mathbf{E}, \tag{5}$$

where $\mathbf{R}^{\star}=[\mathbf{R}_1/\hat{\delta}_1,\ldots,\mathbf{R}_M/\hat{\delta}_M]$ is a variation shared across all scanners, $\mathbf{I}^{\star}=[\mathbf{I}_1^{\star};\ldots;\mathbf{I}_M^{\star}]=[\mathbf{I}_1/\hat{\delta}_1;\ldots;\mathbf{I}_M/\hat{\delta}_M]$ are individual variations shared only in each scanner.

From model (5), $\hat{\mathbf{R}}^{\star}$ and $\hat{\mathbf{I}}^{\star}$ are obtained by

$$\{\hat{\mathbf{R}}^{\star},\hat{\mathbf{I}}^{\star}\}=\underset{\{\mathbf{R}^{\star},\mathbf{I}^{\star}\}}{\arg\min}\left\{\|\Delta-\mathbf{R}^{\star}-\mathbf{I}^{\star}\|_F^2+\lambda\|\mathbf{R}^{\star}\|_{*}+\sum_{i=1}^{M}\lambda_i\|\mathbf{I}_i^{\star}\|_{*}\right\}, \tag{6}$$

where $\|\cdot\|_F^2$ and $\|\cdot\|_{*}$ are the squared Frobenious norm (sum of squared elements) and the nuclear norm (sum of singular values), respectively. The nuclear norm penalties in Equation (6) ensure that the resulting estimates $\hat{\mathbf{R}}^{\star}$, $\hat{\mathbf{I}}^{\star}$ are low-rank (Hastie, Mazumder, Lee, & Zadeh, 2015). Although tuning λ and λ_is may be tricky, we use the recommended values from Park and Lock (2020) by setting $\lambda=\sqrt{p}+\sqrt{n}$ and $\lambda_i=\sqrt{p}+\sqrt{n_i}$, which was shown to perform well with independent Gaussian noise. With λ and λ_is specified, an iterative algorithm can be applied to estimate $\mathbf{R}^{\star}$ and $\mathbf{I}_i^{\star}$s.

In Step (iv), we scale $\hat{\mathbf{R}}^{\star}$ back to $\mathbf{R}^{\star}$ $(\hat{\delta}_i\hat{\mathbf{R}}_i^{\star})$ to make sure $\hat{\mathbf{R}}=\left[\hat{\delta}_1\hat{\mathbf{R}}_1^{\star}\ldots\hat{\delta}_M\hat{\mathbf{R}}_M^{\star}\right]$ is in the original scale. To keep the noise variance homogeneous, we scale $\hat{\mathbf{E}}$ to $\hat{\delta}\hat{\mathbf{E}}$, where $\hat{\delta}^2=\left(\sum_{i=1}^{M}n_i\hat{\delta}_i^2\right)/\left(\sum_{i=1}^{M}n_i\right)$ is the weighted mean of scanner-specified noise variance. Therefore, the final harmonized data is given by

$$\mathbf{Y}^{RELIEF}=\underbrace{\hat{\mathbf{A}}}_{\text{intercepts}}+\underbrace{\hat{\beta}\mathbf{X}'}_{\text{covariate effects}}+\underbrace{[\hat{\delta}_1\hat{\mathbf{R}}_1^{\star};\ldots;\hat{\delta}_M\hat{\mathbf{R}}_M^{\star}]}_{\text{original–scale shared variations}}+\underbrace{\hat{\delta}\hat{\mathbf{E}}}_{\text{rescaled noise}}. \tag{7}$$

2.4. Using covariates in RELIEF

When a primary interest is to test for an association with a covariate of interest, including the covariate in RELIEF may lead to an inflated false positive rate. Intuitively, it is because our objective function (6) does not enforce scores of $\hat{\mathbf{I}}$ to be independent of the covariate of interest. Therefore, we suggest not including covariates of interest when applying RELIEF. In practice, we found that not including any covariates in RELIEF does not result in a noticeable difference because the covariate effects are actually low-rank (with the rank equal to the number of covariates) and are captured by $\mathbf{R}$ (in a high signal-to-noise ratio (SNR)) or by $\mathbf{E}$ (in a low SNR), provided that covariates are independent to scanners. In Section 4, we show that RELIEF still achieves higher power than other harmonization method even when the covariate of interest is not specified as an input in RELEF.

2.5. Preventing distorted covariate effects in RELIEF

Many existing harmonization methods, including Adj Res, ComBat, CovBat, and RELIEF, account for explicit covariate effects in the form of regression, but there might be hidden covariate effects from unobserved covariates. For downstream analyses, it is critical to preserve these effects in the original scale. In RELIEF, such effects correspond to the **R** term, and therefore, we scale $\hat{\mathbf{R}}_i^\star$ back to $\hat{\delta}_i$ in Equation (7) although δ_i were used to characterize variance heterogeneity.

We point out that ComBat (and CovBat that uses ComBat in the first step) models *observed* covariate effects only and all *unobserved* covariate effects are attributed to the residuals. Since residuals are eventually scaled differently for each scanner/site in the harmonization steps, ComBat and CovBat could be prone to distorted covariate effects for unobserved covariates after harmonization, especially when variance heterogeneity across scanners is evident.

3. DATA ANALYSIS

3.1. Data preparation and preprocessing

We used diffusion tensor imaging (DTI) data from Social Processes Initiative in the Neurobiology of the Schizophrenia(s) (SPINS) study to empirically evaluate RELIEF's performance. The study subjects consisted of 256 individuals with schizophrenia spectrum disorders (SSDs) and 175 controls. Subjects were 18–55 years old, and 268 of the participants were males (163 females). Participants with SSDs met DSM-5 diagnostic criteria for schizophrenia, schizoaffective disorder, schizophreniform disorder, delusional disorder, or psychotic disorder not otherwise specified, assessed using the Structured Clinical Interview for DSM (SCID-IV-TR), and had no change in antipsychotic medication or decrement in functioning/support level in the 30 days prior to enrollment. Controls did not have a current or past Axis I psychiatric disorder, except adjustment disorder, phobic disorder, and past major depressive disorder (over 2 years prior; presently unmedicated), or a first-degree relative with a history of psychotic mental disorder. Additional exclusion criteria included a history of head trauma resulting in unconsciousness, a substance use disorder (confirmed by urine toxicology screening), intellectual disability, debilitating or unstable medical illness, or other neurological diseases. Participants also had normal or corrected-to-normal vision. All participants signed an informed consent agreement, and the protocol was approved by the respective research ethics and institutional review boards. All research was conducted in accordance with the Declaration of Helsinki.

The scans were acquired at three different imaging sites, including the Centre for Addiction and Mental Health (CAMH), Maryland Psychiatric Research Center (MPRC), and Zucker Hillside Hospital (ZHH). General Electric 3T MRI scanners were used at CAMH and ZHH (750w Discovery and Signa, respectively), and the Siemens Tim Trio 3T MRI scanner at MPRC. However, during the middle of the study, all study sites switched to Siemens Prisma 3T scanners for data collection. A high-angular resolution axial EPI dual spin echo sequence diffusion scan was acquired on all scanners. Within the limits of scanner hardware, parameters were prospectively harmonized as follows: 60 gradient directions, b = 1,000, 5 b = 0 images, TR = 8,800 ms (one scanner TR = 17,000 ms), TE = 85 ms, FOV = 256 mm; in-plane matrix 128×128, and 2.0 mm isotropic voxels. All images were preprocessed using the same pipeline across sites. Skull-stripping was performed via a two-step process combining FSL (BET) and AFNI to optimize brain extraction, after which MRtrix3 (dwi2mask) was used for brain masking. FSL eddy was used for eddy current-induced distortion and motion correction, including volume-to-volume and within-volume movement (Tournier et al., 2019). Eddy models the effects of participant movement and diffusion eddy currents simultaneously, predicting undistorted data using a Gaussian Process. Eddy also outputs quality control metrics, including average absolute motion (mm) for each participant as one measure of volume-to-volume movement. Fieldmap-free susceptibility distortion correction was performed using BrainSuite (BDP; Bhushan et al., 2015). Outputs were visually inspected after each preprocessing step to ensure data quality.

Participants' white matter tracts were reconstructed using deterministic unscented Kalman Filter (UKF) tractography (Malcolm, Shenton, & Rathi, 2010) in 3D Slicer (https://github.com/SlicerDMRI). The ORG (O'Donnell Research Group) white matter atlas (F. Zhang et al., 2018) was used to parcellate fibers into anatomical tracts. This atlas has been validated across different scanners and protocols (e.g., number of gradient directions, spatial resolutions, b-values; F. Zhang et al., 2019). Metrics were included from 56 deep white matter fiber tracts from the association, cerebellar, commissural, and projection tracts (the cortico-ponto-cerebellar tract was excluded due to parcellation issues), and 16 superficial tract categories according to the brain lobes they connect, resulting in $V = 72$ features. Mean FA values and mean diffusivity (MD) values were calculated along each tract. FA measures the degree to which diffusion of water molecules is restricted by microstructural elements such as cell bodies, axons,

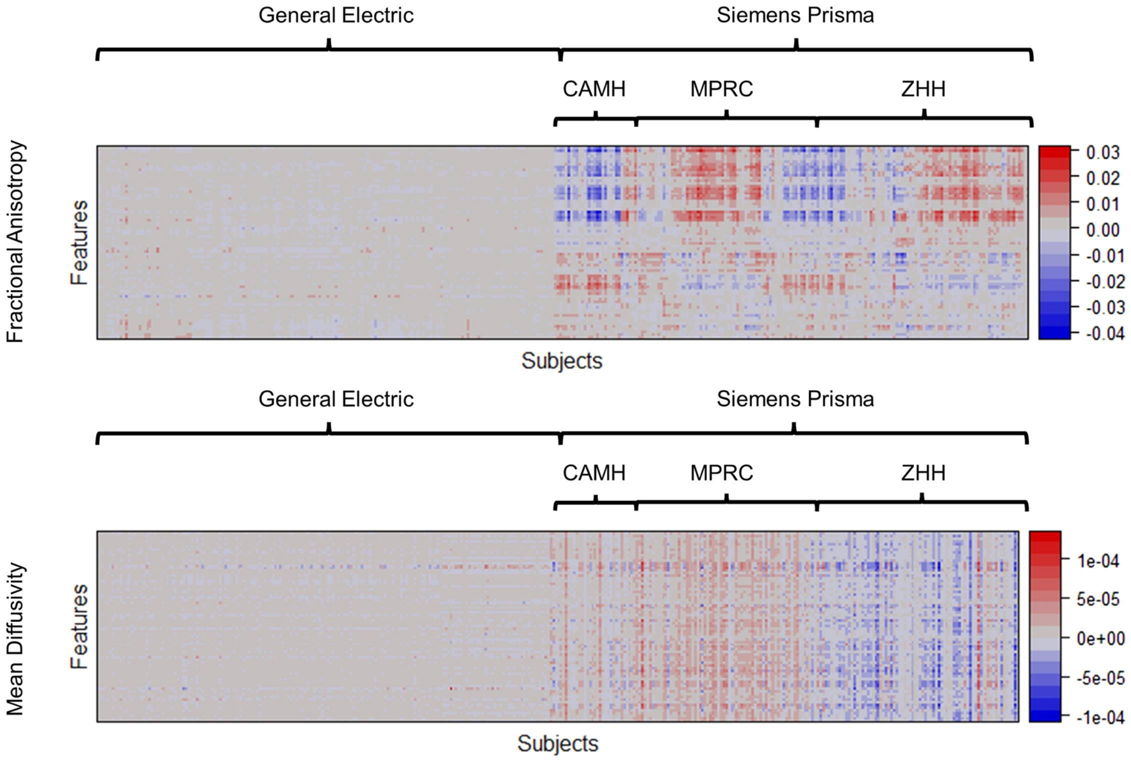

Fig. 2. Heatmaps of the estimated latent scanner-specific variations ($\hat{\mathbf{I}}$) of the FA and MD from the SPINS study. For visualizations, imaging features were reordered by applying hierarchical clustering; subjects scanned by General Electric 3T were reordered separately, and subjects scanned by Siemens Prisma 3T were reordered within each site subgroup (CAMH, MPRC, and ZHH). Feature indices were also reordered by applying hierarchical clustering to concatenated $\mathbf{I}$ for FA and MD. RELIEF identified substantial variations present mostly on Siemens Prisma but not on General Electric, and the variations are highly associated with sites.

myelin, and other constituents of cytoskeleton (Beaulieu, 2002). MD is a measure of the magnitude of water diffusion, independent of direction (O'Donnell & Westin, 2011). Visual quality control was performed after initial tractography, registration to the ORG atlas, and tract creation. Data from seven participants were excluded on the basis of missing or poor tractography for >15 tracts across the whole brain.

Since the number of samples from Siemens Tim Trio is small, we used images from two scanner types (GE and SP) in our analysis. Participants without DTI data were also excluded from the study. The final sample consists of 351 subjects across 2 scanner types, with 172 subjects imaged on scanners manufactured by GE (67 females, 111 patients, age 18-55), 179 on Prisma scanners manufactured by Siemens (71 females, 98 patients, age 18-55).

3.2. Results

We harmonized data by using RELIEF, ComBat, CovBat, and AdjRes. We used age, age^2, gender, diagnosis, an interaction between age and gender (age × gender), and an interaction between age and diagnosis (age × diagnosis) to model covariate effects in harmonization.

Figure 2 shows the heatmap of the estimated latent scanner effects $\hat{\mathbf{I}}$ of RELIEF for the FA and MD data from the SPINS study. As RELIEF's crucial components, the latent scanner effects are identified and removed to reduce the inter-scanner variations directly. In Figure 2, the most scanner-specific variations were attributed to Siemens Prisma for both FA and MD. To investigate the potential sources of latent Siemens Prisma-specific variations in relation to existing non-biological information, we applied hierarchical clustering to the site subgroups of $\hat{\mathbf{I}}_{SP}$ in Figure 2 and reordered subjects within Siemens Prisma so that $\hat{\mathbf{I}}_{SP}$ within the same site were arranged together. We observed the latent scanner effects within each site tended to share similar patterns, which suggests that the variations in $\hat{\mathbf{I}}_{SP}$ are highly associated with sites.

We performed statistical analysis to quantify the relationship between existing non-biological information, including site information and motion parameters. In

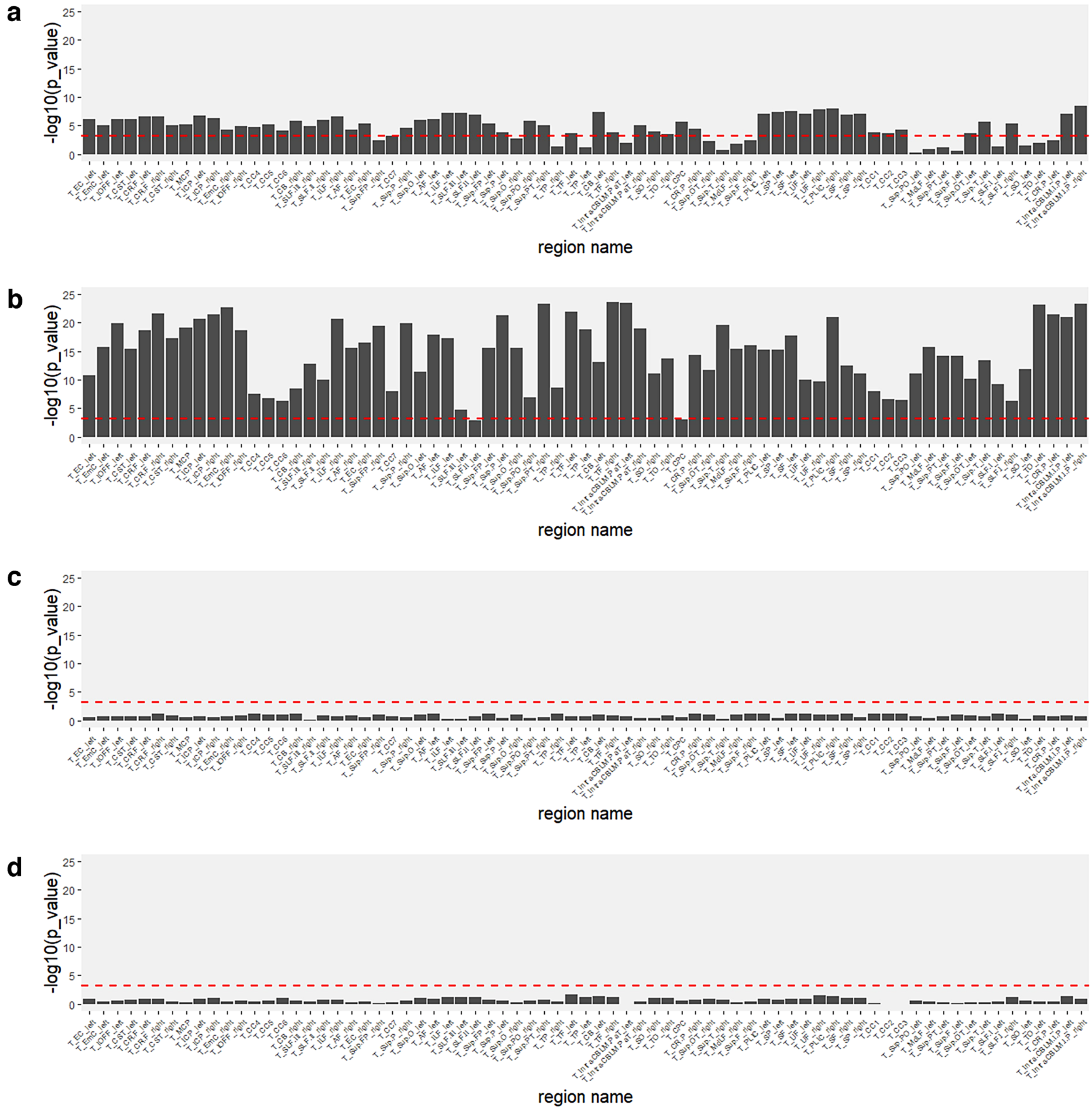

Fig. 3. Investigating the potential sources of latent Siemens Prisma-specific variations in relation to existing non-biological information (site information and a motion parameter). (a) and (b) show one-way ANOVA p values of (a) FA and (b) MD in relation to three study sites. (c) and (d) show p values for the correlation between latent factors and the average absolute motion from the reference volume (in mm) for (c) FA and (d) MD. All p values were negative log-transformed (with base 10) for visualizations. The red dashed horizontal line is Bonferroni-corrected threshold ($0.05/72 \approx 6.9 \times 10^{-4}$). The region names agree with the order in Figure 2.

Figure 3 (a) and (b), we performed one-way ANOVA to compare different latent scanner effects of the Siemens Prisma across sites for FA and MD data, respectively. We found that the latent factors of most features specific to Simens Prisma were highly associated with the sites, particularly for MD data. In Figure 3 (c) and (d), we performed correlation tests between the latent scanner effects in the Siemens Prisma scanner and the motion

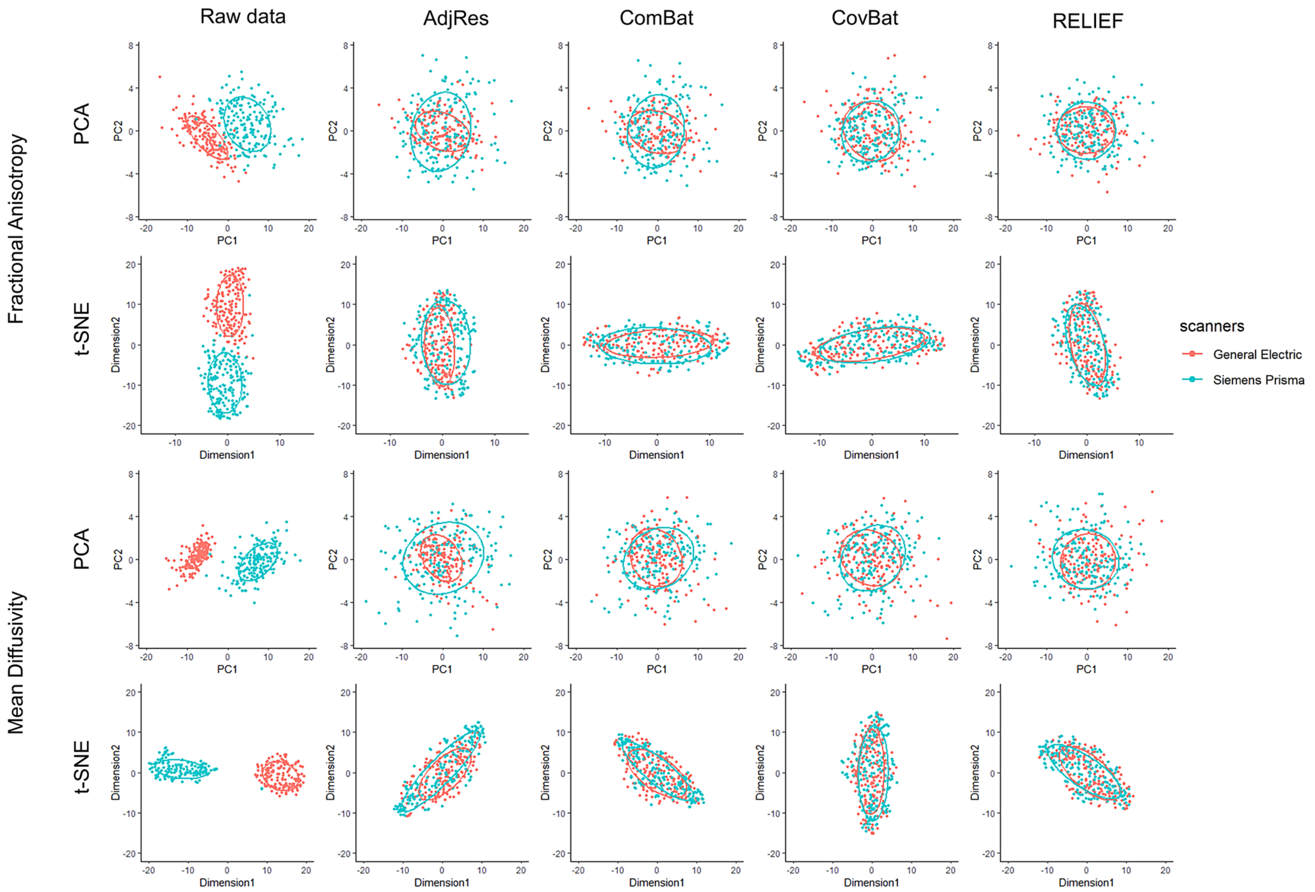

Fig. 4. Scatterplots of principal component scores and t-SNE scores before and after applying harmonization to the SPINS DTI data.

parameter for FA and MD data, respectively. We calculated the average absolute motion from the reference volume (in mm) to represent subject motion during the scan and averaged it for the six motion parameters (three translations and three rotations). Our findings revealed that the latent factors showed no significant associations with the motion parameter. Overall, our analyses provided insights into how existing non-biological information can impact the interpretation of latent scanner-specific variations.

To visualize whether most variations in the data are still associated with scanners after harmonization, we applied two unsupervised data reduction techniques: principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) to the original and harmonized FA and MD data from diffusion tensor imaging (DTI). As a nonlinear technique, t-SNE emphasizes preserving the variations in the local structure of the data, while PCA focuses more on preserving variations in the overall data set. The data projected into the first two PCs/dimensions are presented in Figure 4. For raw data, we observed that most variations are clearly explained by the scanner information (General Electric vs. Siemens Prisma). For AdjRes and ComBat, despite evidences of higher data quality, there is heteroscedasticity of ellipses across scanners, which indicates that there are still unremoved latent scanner effects. For CovBat and RELIEF, both PC scores and t-SNE scores appear to be distributed similarly across scanners, which suggests the variations associated with scanners are substantially removed.

To evaluate if scanner-specific latent patterns are well-removed, we computed the empirical covariances by scanners as well as the difference between two scanner-specific covariances. Figure 5 shows that the covariance differences remain notable in AdjRes harmonized data. ComBat and CovBat performed slightly better than AdjRes in mitigating covariance scanner effects. Notably, however, these covariance differences are considerably reduced with RELIEF. We also quantified these differences

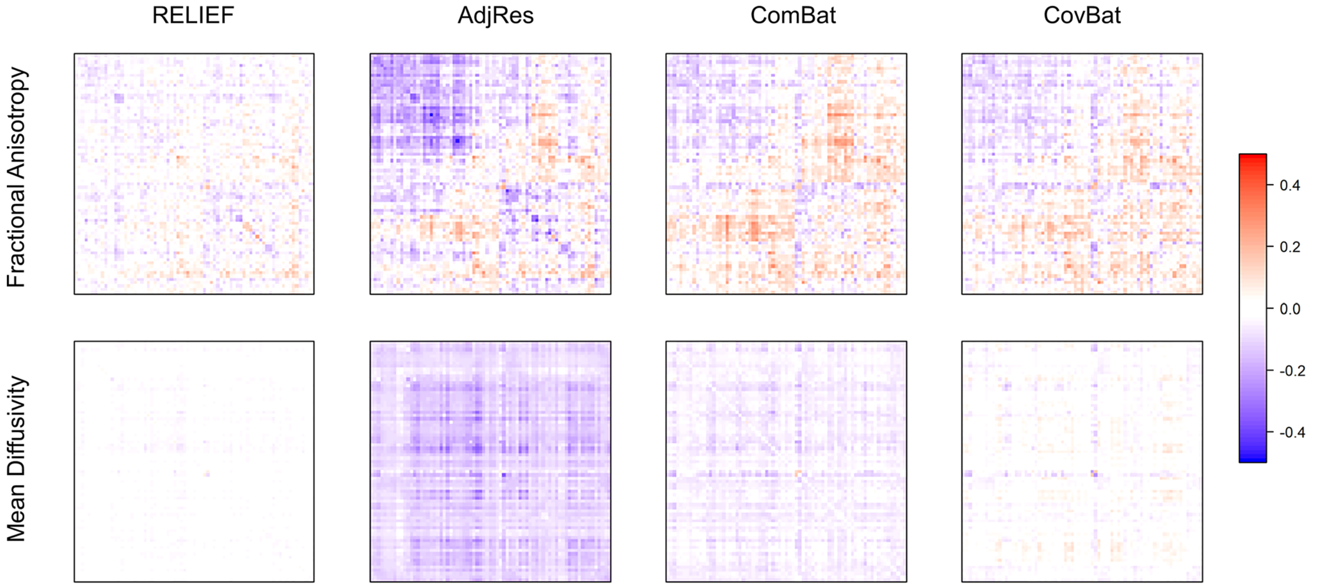

Fig. 5. The difference of scanner-specific covariance matrices for harmonized SPINS data (GE–SP). The order of the features agrees with Figures 2 and 3. The x-axis and y-axis indicate regions of interest, which are explicitly illustrated in the x-axis of Figure 3. The color bar shows the range of values of the differences in covariances. RELIEF reveals the lowest difference between two covariances.

in covariances by the Frobenius norm of the scanner-specific covariance matrices. For FA, the norm for RELIEF was the lowest (**3.70**) followed by CovBat (5.77), ComBat (6.19), and AdjRes (8.10). For MD, the norm for RELIEF was also the lowest (**1.45**) followed by CovBat (2.29), ComBat (4.15), and AdjRes (8.74). These results suggest the superior performance of RELIEF in constructing homogeneous covariances.

We also used Quadratic Discriminant Analysis (QDA) to evaluate how data harmonized using each approach predicts scanners. A harmonization method that performs *better* in removing scanner effects would result in *worse* predictive performance. Using machine-learning methods to predict scanners from harmonized data has been adopted in previous work in evaluating the performance of different harmonization methods (Chen, Beer, et al., 2022; J.-P. Fortin et al., 2018). We chose QDA because the classifier is constructed based on the mean vectors and covariance matrices only, where differences in predictive performances are attributed to the harmonization of scanner-specific means and covariances. Using leave-one-out cross-validation, we computed the average accuracy, ROC curve, and its area under the curve (AUC) for each harmonized data after regressing out covariate effects. For FA, the RELIEF method achieved the lowest prediction accuracy (**49.6%**) close to a random prediction, followed by CovBat (59.3%), ComBat (66.1%), and AdjRes (70.1%). For MD, RELIEF also achieved the lowest prediction accuracy (**61.0%**) followed by CovBat (82.6%), ComBat (83.2%), and AdjRes (87.5%) The results of the AUC, shown in Figure 6, were similar to the prediction accuracy, suggesting the lowest AUC for RELIEF.

Lastly, we investigated whether RELIEF preserves the biological variability in the data. This step is necessary because the multivariate harmonization methods could be prone to potentially overkilling too much variation, including biological variations. Here, we evaluated whether the different harmonization methods maintain the biological associations of interest through multiple linear regression. For each FA/MD feature in each harmonized data, we built a regression for each feature by using the same set of covariates (age, age^2, gender, diagnosis, age × gender, and age × diagnosis) as the harmonization step. We then computed *t* statistics of the estimated coefficients across all covariates and features. The boxplots of *t* statistics are shown in Figure 7. We observed that, for FA data, the magnitude of *t* statistics of all harmonized data appeared to be similar, which confirms that RELIEF did not lose biological information compared with other methods. However, for MD data, RELIEF clearly showed more significant associations with diagnosis and age × diagnosis than other methods, which suggests that RELIEF not only provided a thorough removal of scanner effects but also maintained biological associations well.

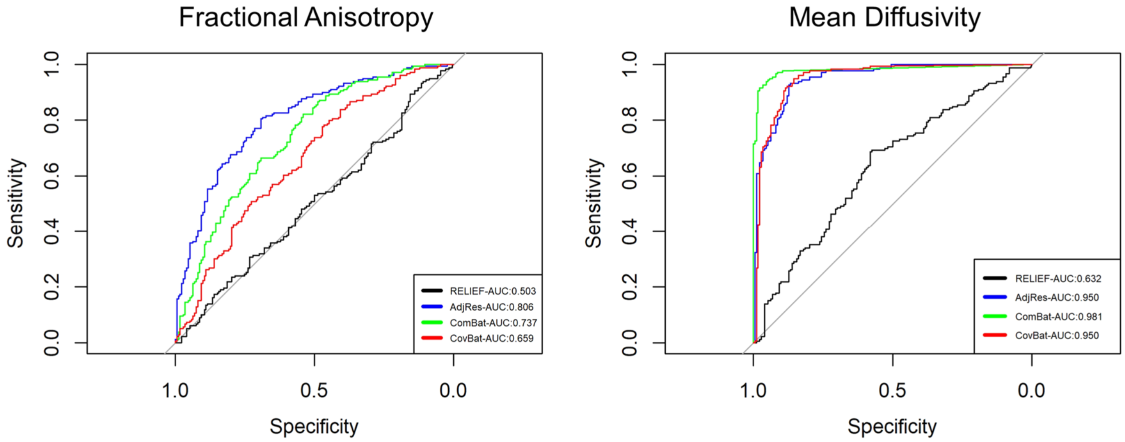

Fig. 6. The ROC curves for predicting scanners by using SPINS data harmonized by different methods. We used QDA as a classifier and leave-one-out cross-validation (LOOCV) to obtain individualized predictions. The ROC curve of the RELIEF was closest to the diagonal line, suggesting that it successfully harmonized latent inter-scanner biases.

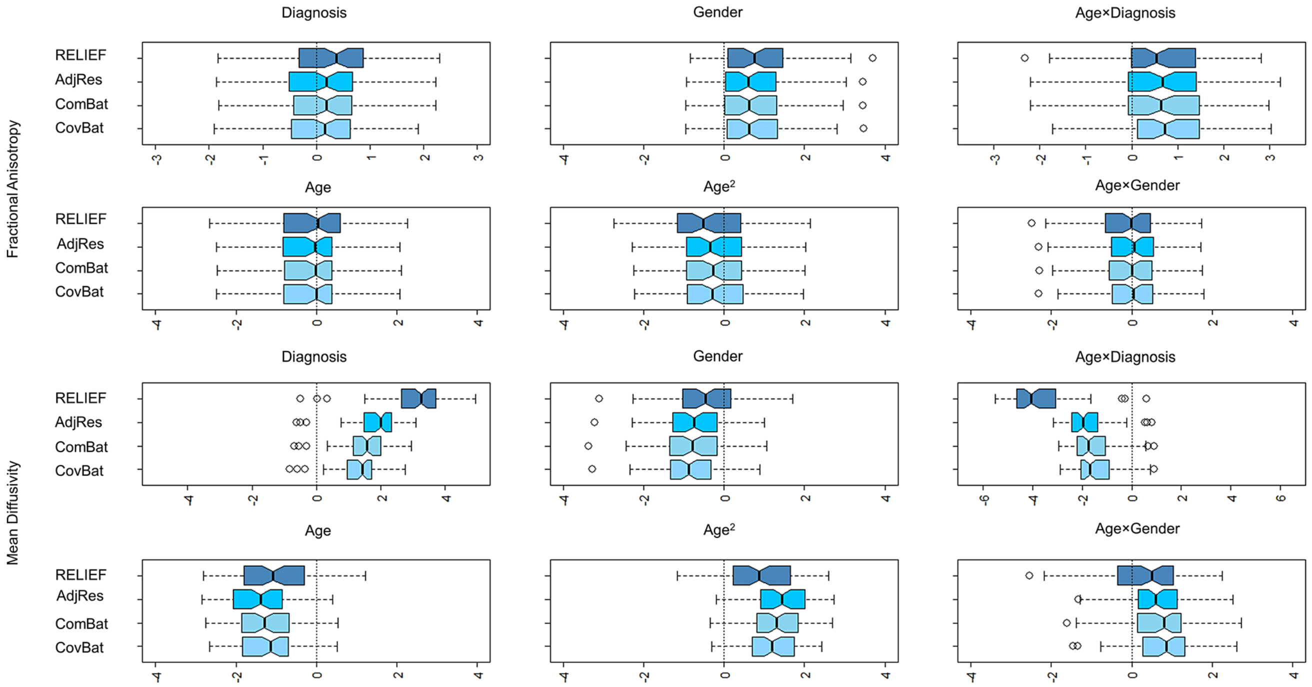

Fig. 7. The boxplots for *t* statistics for each biological covariate used in our analysis.

4. SIMULATION STUDIES

4.1. Simulation designs

In this section, we performed extensive simulation studies to evaluate the performance of RELIEF and to compare it to other methods in controlled settings. We included ComBat, CovBat, and AdjRes as our competitors and evaluated how well-harmonized data preserve biological variations through power analysis. To evaluate the control of false positives and power, we used two models to generate heterogeneous covariances across scanners.

4.1.1. *Simulation 1: RELIEF model*

We generated data using the sum of low-rank features following Equation (4). We simulated 1,000 null data sets with $n_1 = n_2 = 50$ (so that $n = 100$), and $V = 100$ features. Our data-generating model is summarized by

$$\underbrace{\mathbf{Y}}_{\text{rank }100} = \underbrace{\mathbf{A}}_{\text{rank}1} + \underbrace{\beta\mathbf{X}'}_{\text{rank}4} + \underbrace{\mathbf{\Gamma}}_{\text{rank}2} + \underbrace{\mathbf{R}}_{\text{rank }3} + \underbrace{c\cdot\mathbf{I}}_{\text{rank }6} + \left[\delta_1\mathbf{E}_1,\ldots,\delta_M\mathbf{E}_M\right].$$

We used four nuisance covariates for the covariate effects, where each element of β and each row of $\mathbf{A}$ were generated from $\mathcal{N}(0,1^2)$. The covariate vector for each subject was generated from the multivariate normal distribution with zero means, and we used AR(1) for the covariance matrix with the autocorrelation parameter 0.2. Second, we generated $\mathbf{R}$ by first generating a $V \times n$ matrix whose entries are drawn from $\mathcal{N}(0,1^2)$, then taking the first three principal components. Similarly, we generated each $\mathbf{I}_i$ by generating a $V \times n_i$ matrix using $\mathcal{N}(0,1^2)$ then taking the top 3 principal components. Lastly, we also generated the additive scanner effect (location) γ_{iv} by fixing it to be the same for all i and from $\mathcal{N}(0,\ 1.5^2)$, and multiplicative scanner effect (scale) δ_i from Uniform (1, 1.5). Finally, the elements of E were generated from $\mathcal{N}(0,1^2)$.

The constant c was chosen between 0, 1, 2, 3 to evaluate the impact of scanner-specific latent patterns on statistical power. Note that we also considered $c = 0$ to investigate whether it has comparable performance when the data-generating model does not include latent scanner effects.

4.1.2. *Simulation 2: CovBat model*

We generated data by modifying the simulation design introduced by Chen, Beer, et al. (2022). To address potential covariance scanner effects, CovBat model uses principal component (PC) scores to shift each within-scanner covariance to the pooled covariance structure. Therefore, the design aimed to evaluate whether harmonization methods can approximate the underlying covariance structure when covariance scanner effects are captured by its PC shifts.

We simulated 1,000 null data sets based on SPINS data so that $n_1 = 172, n_2 = 179$ (so that $n = 351$) and $V = 72$ features. The data y_{ijv} was generated by $y_{ijv} = \alpha_v + \gamma_{iv} + \delta_{iv}\varepsilon_{ijv}$, where $\alpha = (\alpha_1,\ldots,\alpha_V)'$ is the sample mean vector of Scanner General Electric observations in the SPINS data. The additive scanner effects $\gamma_i = (\gamma_{i1},\ldots,\gamma_{iV})'$'s are vectors drawn from $\mathcal{N}(0,0.1^2)$. For multiplicative scanner effects, we used $\delta_{1v} \sim \mathcal{IG}(46,\ 50)$ and $\delta_{2v} \sim \mathcal{IG}(51,\ 50)$ following Chen, Beer, et al. (2022). From the sample correlation matrix of DTI-FA observations in the SPINS data (termed $\mathbf{S}$) with its corresponding eigen decomposition $\mathbf{S} = \sum_{l=1}^{72} \hat{\lambda}_l \hat{\psi}_l \hat{\psi}_l'$, we generated $\epsilon_{ij} = (\epsilon_{ij1},\ldots,\epsilon_{ijV})'$ that contained scanner-specific shifts. The design was to investigate how the rank of the covariance effect influences harmonization results, and we generated error terms by $\epsilon_{ij} \sim \mathcal{MVN}\left(\mathbf{0}, \mathbf{S} + c_i \sum_{l=1}^{L} \hat{\lambda}_l \hat{\psi}_l \hat{\psi}_l'\right)$, where $c_1 = -\frac{3}{4}$ and $c_2 = \frac{3}{4}$. We considered different L including $L = 0, 10, 20, 30$.

In both simulation designs, we generated our covariate of interest, Z_k $(k = 1,\ldots,n)$, randomly from 0 or 1, for evaluation of power. We randomly chose 20% (for Simulation 1) and 50% (for Simulation 2) of features and added $\tau_v \cdot Z_k$ to the null data, where $\tau_v \geq 0$ is the effect size for the vth feature, which controls whether the simulated data follow the null hypothesis $H_0 : \tau_1 = \ldots = \tau_V = 0$ or the alternative hypothesis H_1 : at least one of $\tau_v \neq 0$ $(v = 1,\ldots,V)$. We used permutation to control family-wise error rate (FWER) at 5%.

4.2. Simulation results

The results for Simulation 1 are summarized in the first row of Figure 8. RELIEF controlled family-wise error properly, with empirical FWER of 0.044, 0.047, 0.048, and 0.052 regardless of the choice of c. In our simulations, while other methods controlled FWER appropriately in most scenarios, CovBat was conservative in controlling false positives when the proportion of individual latent patterns increased. In terms of power, RELIEF's performance was nearly the same as ComBat or CovBat even when there are no latent scanner effects (i.e., $c = 0$), which supports the robustness of the proposed method. Also, as the degree of latent scanner effects (c) increased, RELIEF showed substantial power gain compared to others, partially because it correctly identified and removed the scanner-specific latent patterns in the data. The lower power of ComBat and AdjRes is expected as they do not consider these latent patterns in their model, and the lower power of CovBat is also expected because RELIEF's data-generating model is different from CovBat's assumption on PC shifts.

The results for Simulation 2 are summarized in the second row of Figure 8. RELIEF's empirical FWERs are 0.05, 0.051, 0.038, and 0.052 for $L = 0,\ 10,\ 20,\ 30$, while ComBat, CovBat are conservative in controlling false positives when covariance scanner effects exist. For power, we note that when covariance scanner effects do not exist ($L = 0$), all harmonization methods increased statistical power and performed similarly, except for AdjRes whose power was lower. When L was large, RELIEF still showed superior performance to other meth-

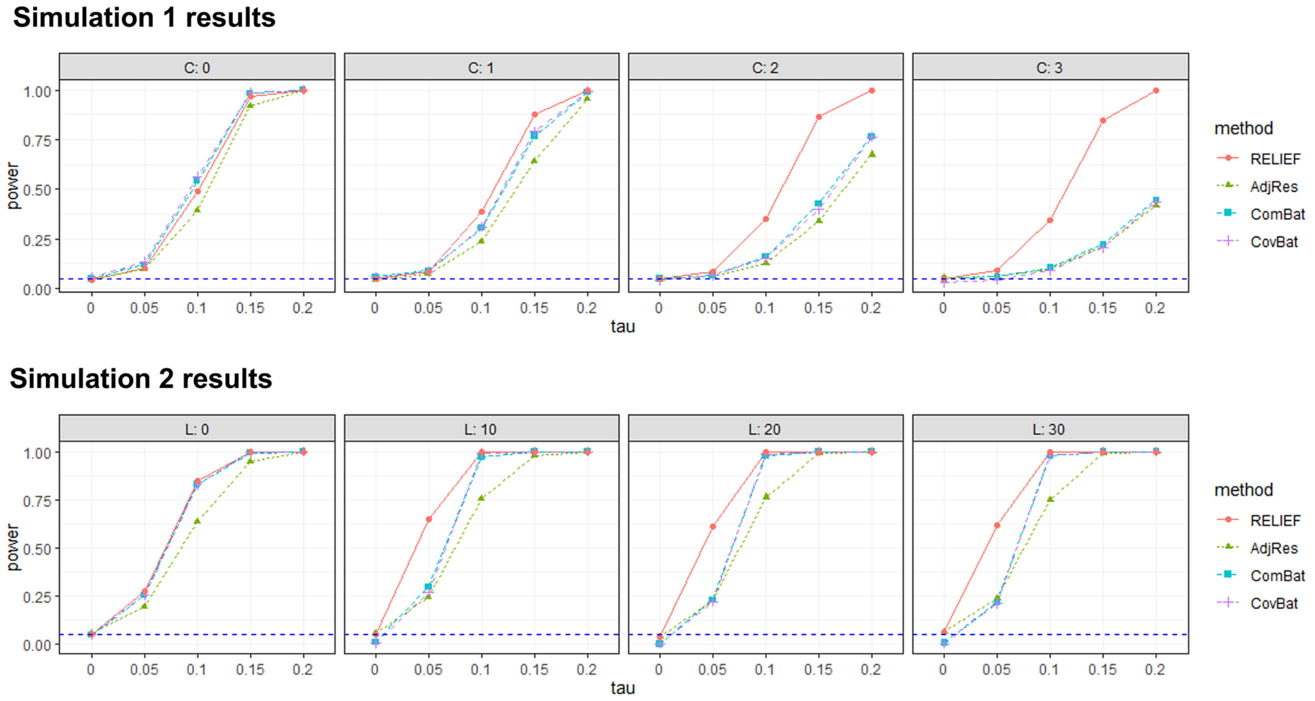

Fig. 8. Summary of power for four harmonization methods. From the first row, the plots from left to right are with the increased proportion of individual latent patterns. From the second row, the plots from left to right are with the increased rank of the covariance effect. The blue dashed horizontal line is FWER = 0.05. RELIEF controls for false positives accurately and shows superior power to competitors in both settings.

ods, which supports the robustness of RELIEF even when the data-generating model did not follow the assumption of RELIEF. In addition, when SNR was low (i.e., $\tau = 0.05$), RELIEF gained higher power than competitors, which supports its ability to denoise scanner effects and preserve true biological associations.

4.3. Additional simulations

To address Section 2.4 empirically, we repeated Simulation 1 to evaluate FWER when the covariate of interest was specified in RELEF. In this simulation, we obtained empirical FWER values of 0.058 (95% CI: (0.052, 0.064)) when $c = 0$ and 0.159 (95% CI: (0.152, 0.166)) when $c = 1$, indicating that RELIEF, when covariates of interest were specified in the model, had inflated false positives.

5. DISCUSSION

We proposed a novel harmonization method, called RELIEF, that estimates and removes both explicit (additive and multiplicative) and latent scanner effects. RELIEF aligns with ongoing efforts to integrate neuroimaging data collected from different scanners or sites. In particular, our methods address covariance heterogeneity across different scanners, which has been a promising direction in mitigating inter-scanner biases. Our approach provides an interpretable way to harmonize heterogeneous covariances by modeling scanner-specific latent patterns under the low-rank assumption. We characterized inter-scanner bias with (i) scanner-specific means (locations), (ii) scanner-specific variances (scales), and (iii) scanner-specific latent patterns. We showed that identification of (iii), which has been overlooked in previous methods, is critical in homogenizing data from multi-site, multi-scanner neuroimaging studies.

RELIEF is a general multivariate approach that does not impose data-specific assumptions. It also does not require traveling subjects or matched controls that are often needed in supervised harmonization methods, which are infeasible in many imaging studies. Also, as we extend a regression-based approach, preserving clinical covariate effects is straightforward. Moreover, it also preserves shared variations from unobserved covariates or non-linear covariate effects using a low-dimensional representation of such variations, in which existing regression-based harmonization methods are limited.

In the analysis of the fractional anisotropy (FA) and mean diffusivity (MD) data from the SPINS study, where study samples were scanned using General Electric or Siemens Prisma scanners, we showed that there are substantial variations specific to Siemens Prisma. Notably, our data analysis reveals that these latent scanner effects for Siemens Prisma are heterogeneous across features (Fig. 2). This result aligns with previous studies showing that inter-site variability in fractional anisotropy is specific to tissues or regions (J.-P. Fortin et al., 2017; Vollmar et al., 2010). RELIEF, which removed these variations in addition to the scanner-specific means and variance, successfully impaired the detection of scanners with a machine-learning method, resulting in a more homogeneous covariance as expected. A correlation analysis with existing non-biological information helped us understand the mechanism that induces these latent scanner effects.

RELIEF is not without limitations. First, our current approach is evaluated with a moderate number of samples. RELIEF assumes that the original data matrix consists of low-rank signals (including latent scanner effects) plus full-rank noises to scale data and choose tuning parameters. To detect these low-rank variations well, it requires a moderate number of samples to ensure the objective function of RELIEF performs more promisingly than simplified methods (e.g., ComBat) with fewer assumptions. Second, although low-rank decomposition is a useful way to capture arbitrary covariance structures, it might not always be the case when there is structured covariance in imaging data. For example, vertex-level cortical thickness data has at most 160,000 features in each brain hemisphere in FreeSurfer and reveals a high degree of spatial autocorrelation. In such a case, the low-rank assumption made in RELIEF should be evaluated carefully (Karayumak et al., 2019; Mirzaalian et al., 2016). Also, although RELIEF does not require intense cross-validation to choose tuning parameters or ranks, it requires applying singular value decomposition (SVD) iteratively, and the computational cost increases non-linearly with increased sample size (n) or features (V). Therefore, it takes more time than existing methods (e.g., ComBat), whose computation time increase linearly with V. However, the computation time for RELIEF is still moderate in most downstream neuroimaging data analyses with, at most, up to hundreds of features. More importantly, we believe the powerful performance of RELIEF outweighs the cost of some additional computation time.

Also, there were recent investigations showing how pre-processing can affect the performance of ComBat harmonization, which could also be the case in RELIEF. Cetin-Karayumak et al. (2020) evaluated the effect of minor differences in pre-processing on ComBat's performance for harmonization of fractional anisotropy (FA) data across sites and showed that minor differences in the preprocessing steps resulted in non-linear changes in the input data. Because the SPINS study performed consistent preprocessing pipelines across sites, we expect its impact on our analysis to be marginal. Still, evaluating the robustness of RELIEF with respect to different preprocessing pipelines would be an interesting area of research, which we leave as future work.

RELIEF is the first approach that adopted the structured factorization of interlinked matrices into the data harmonization context, which used the concept of latent variables to characterize scanner effects. In the past decade, there have been a number of methodological developments in linked matrix factorization (Feng, Jiang, Hannig, & Marron, 2018; Gaynanova & Li, 2019; Lock et al., 2013), which provided novel insights into understanding multimodal data (Q. Yu et al., 2017), disease subtypes, or clustering. We believe more methodological research on data harmonization from the viewpoint of the linked matrix factorization would lead to further improvements in the harmonization quality.

To summarize, we proposed a new harmonization method, RELIEF, that contributes to ongoing efforts on integrating heterogeneous multi-site, multi-scanner studies in neuroimaging. Our novel contribution is the development of a multivariate harmonization method that captures scanner-specific latent factors, which have not been addressed in existing methods. With the three-source characterization of inter-scanner biases (location, scale, latent), RELIEF shows promising results in harmonizing all of them, eventually resulting in higher power in association studies than existing harmonization methods.

6. SOFTWARE

RELIEF is made publicly available as an R package on GitHub: https://github.com/junjypark/RELIEF. It requires the same input as neuroComBat (https://github.com/Jfortin1/ComBatHarmonization) (imaging data matrix, covariates, and scanner information), producing harmonized imaging data in the same format. Our harmonization took approximately 4 seconds on a Macbook Pro 2018 to harmonize data with 72 imaging features from 351 subjects, which supports the computational efficiency of the proposed method.

DATA AND CODE AVAILABILITY

The R package for implementing RELIEF is publicly available at https://github.com/junjypark/RELIEF.

AUTHOR CONTRIBUTIONS

Rongqian Zhang: Methodology, Software, Formal analysis, Validation, Visualization, and Writing—original draft. Lindsay D. Oliver: Data curation, Supervision, and Writing—review & editing. Aristotle N. Voineskos: Supervision, Funding acquisition, and Writing—review & editing. Jun Young Park: Conceptualization, Methodology, Validation, Supervision, Software, Writing—review & editing, and Funding acquisition.

DECLARATION OF COMPETING INTEREST

None.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful comments. The authors would like to thank all participants for their contribution to this work, and the research staff who performed data collection and management. The SPINS study was supported by the National Institute of Mental Health (1/3R01MH102324-01, 2/3R01MH102313-01, 3/3R01MH102318-01). L.D.O. was supported by the Brain & Behavior Research Foundation. A.N.V. was supported by the National Institute of Mental Health (1/3R01MH102324 & 1/5R01MH114970), Canadian Institutes of Health Research, Canada Foundation for Innovation, CAMH Foundation, and University of Toronto. J.Y.P. was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2022-04831) and the University of Toronto's Data Science Institute (Catalyst Grant). The computing resources were enabled in part by support provided by the University of Toronto and the Digital Research Alliance of Canada (alliancecan.ca).

REFERENCES

Beaulieu, C. (2002). The basis of anisotropic water diffusion in the nervous system—A technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, *15*(7–8), 435–455. https://doi.org/10.1002/nbm.782

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., … Initiative, A. D. N. (2020). Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, *220*, 117129. https://doi.org/10.1016/j.neuroimage.2020.117129

Bhushan, C., Haldar, J. P., Choi, S., Joshi, A. A., Shattuck, D. W., & Leahy, R. M. (2015). Co-registration and distortion correction of diffusion and anatomical images based on inverse contrast normalization. *Neuroimage*, *115*, 269–280. https://doi.org/10.1016/j.neuroimage.2015.03.050

Cetin-Karayumak, S., Stegmayer, K., Walther, S., Szeszko, P. R., Crow, T., James, A., … Rathi, Y. (2020). Exploring the limits of ComBat method for multi-site diffusion MRI harmonization. *bioRxiv*, 2020–2011. https://doi.org/10.1101/2020.11.20.390120

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara R. T., Shou, H., & Initiative, A. D. N. (2022). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping*, *43*(4), 1179–1195. https://doi.org/10.1002/hbm.25688

Chen, A. A., Luo, C., Chen, Y., Shinohara, R. T., Shou, H., Initiative, A. D. N., et al. (2022). Privacy-preserving harmonization via distributed ComBat. *NeuroImage*, *248*, 118822. https://doi.org/10.1016/j.neuroimage.2021.118822

Chen, A. A., Srinivasan, D., Pomponio, R., Fan, Y., Nasrallah, I. M., Resnick, S. M., … Shou, H. (2022). Harmonizing functional connectivity reduces scanner effects in community detection. *NeuroImage*, *256*, 119198. https://doi.org/10.1016/j.neuroimage.2022.119198

Dansereau, C., Benhajali, Y., Risterucci, C., Pich, E. M., Orban, P., Arnold, D., & Bellec, P. (2017). Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. *Neuroimage*, *149*, 220–232. https://doi.org/10.1016/j.neuroimage.2017.01.072

Feng, Q., Jiang, M., Hannig, J., & Marron, J. (2018). Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, *166*, 241–265. https://doi.org/10.1016/j.jmva.2018.03.008

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., … Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., … Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., Initiative, A. D. N., et al. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, *132*, 198–212. https://doi.org/10.1016/j.neuroimage.2016.02.036

Gavish, M., & Donoho, D. L. (2017). Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, *63*(4), 2137–2152. https://doi.org/10.1109/TIT.2017.2653801

Gaynanova, I., & Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics*, *75*(4), 1121–1132. https://doi.org/10.1111/biom.13108

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., … Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *Neuroimage*, *32*(1), 180–194. https://doi.org/10.1016/j.neuroimage.2006.02.051

Hastie, T., Mazumder, R., Lee, J. D., & Zadeh, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research*, *16*(1), 3367–3402. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6530939/

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Karayumak, S. C., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., … Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. *Neuroimage*, *184*, 180–200. https://doi.org/10.1016/j.neuroimage.2018.08.073

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, *3*(9), e161. https://doi.org/10.1371/journal.pgen.0030161

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, *7*(1), 523. https://doi.org/10.1214/12-aoas597

Lock, E. F., Park, J. Y., & Hoadley, K. A. (2022). Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. *The Annals of Applied Statistics*, *16*(1), 193. https://doi.org/10.1214/21-AOAS1495

Malcolm, J. G., Shenton, M. E., & Rathi, Y. (2010). Filtered multitensor tractography. *IEEE Transactions on Medical Imaging*, *29*(9), 1664–1675. https://doi.org/10.1109/TMI.2010.2048121

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., … Rathi, Y. (2016). Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage*, *135*, 311–323. https://doi.org/10.1016/j.neuroimage.2016.04.041

O'Donnell, L. J., & Westin, C.-F. (2011). An introduction to diffusion tensor image analysis. *Neurosurgery Clinics*, *22*(2), 185–196. https://doi.org/10.1016%2Fj.nec.2010.12.004

Park, J. Y., & Lock, E. F. (2020). Integrative factorization of bidimensionally linked matrices. *Biometrics*, *76*(1), 61–74. https://doi.org/10.1111/biom.13141

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., … Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, *208*, 116450. https://doi.org/10.1016/j.neuroimage.2019.116450

Takao, H., Hayashi, N., & Ohtomo, K. (2014). Effects of study design in multi-scanner voxel-based morphometry studies. *Neuroimage*, *84*, 133–140. https://doi.org/10.1016/j.neuroimage.2013.08.046

Tournier, J.-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., … Connelly, A. (2019). Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage*, *202*, 116137. https://doi.org/10.1016/j.neuroimage.2019.116137

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., … Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners. *Neuroimage*, *51*(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., … Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, *39*(11), 4213–4227. https://doi.org/10.1002/hbm.24241

Yu, Q., Risk, B. B., Zhang, K., & Marron, J. (2017). JIVE integration of imaging and behavioral data. *NeuroImage*, *152*, 38–49. https://doi.org/10.1016/j.neuroimage.2017.02.072

Zhang, F., Wu, Y., Norton, I., Rathi, Y., Golby, A. J., & O'Donnell, L. J. (2019). Test–retest reproducibility of white matter parcellation using diffusion mri tractography fiber clustering. *Human Brain Mapping*, *40*(10), 3041–3057. https://doi.org/10.1002/hbm.24579

Zhang, F., Wu, Y., Norton, I., Rigolo, L., Rathi, Y., Makris, N., & O'Donnell, L. J. (2018). An anatomically curated fiber clustering white matter atlas for consistent white matter tract parcellation across the lifespan. *NeuroImage*, *179*, 429–447. https://doi.org/10.1016/j.neuroimage.2018.06.027

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, *2*(3), lqaa078. https://doi.org/10.1093/nargab/lqaa078

Zhu, T., Hu, R., Qiu, X., Taylor, M., Tso, Y., Yiannoutsos, C., … Zhong, J. (2011). Quantification of accuracy and precision of multi-center DTI measurements: A diffusion phantom and human brain study. *Neuroimage*, *56*(3), 1398–1411. https://doi.org/10.1016/j.neuroimage.2011.02.010