SAN: mitigating spatial covariance heterogeneity in cortical thickness data collected from multiple scanners or sites

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

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

Abstract

In neuroimaging studies, combining data collected from multiple study sites or scanners is becoming common to increase the reproducibility of scientific discoveries. At the same time, unwanted variations arise by using different scanners (inter-scanner biases), which need to be corrected before downstream analyses to facilitate replicable research and prevent spurious findings. While statistical harmonization methods such as ComBat have become popular in mitigating inter-scanner biases in neuroimaging, recent methodological advances have shown that harmonizing heterogeneous covariances results in higher data quality. In vertex-level cortical thickness data, heterogeneity in spatial autocorrelation is a critical factor that affects covariance heterogeneity. Our work proposes a new statistical harmonization method called SAN (Spatial Autocorrelation Normalization) that preserves homogeneous covariance vertex-level cortical thickness data across different scanners. We use an explicit Gaussian process to characterize scanner-invariant and scanner-specific variations to reconstruct spatially homogeneous data across scanners. SAN is computationally feasible, and it easily allows the integration of existing harmonization methods. We demonstrate the utility of the proposed method using cortical thickness data from the Social Processes Initiative in the Neurobiology of the Schizophrenia(s) (SPINS) study. SAN is publicly available as an R package.

Keywords: cortical thickness; covariance heterogeneity; Gaussian process; inter-scanner biases; neuroimaging

*Corresponding author
Email address: junjy.park@utoronto.ca (Jun Young Park)

Preprint submitted to a journal

1. Introduction

Emerging techniques facilitate the quantification of human cerebral cortex properties through magnetic resonance imaging (MRI), such as cortical thickness, surface area, and gyrification [1]. These advancements have profound implications in the study of brain structures and functions. Cortical thickness is defined as the spatial span between the gray matter and white matter surfaces of the cerebral cortex, for which subtle variations in brain structure may enhance the understanding of neurological disorders, developmental changes, and potential cognitive processes. Notably, alterations in cortical thickness have been implicated in normal aging [2, 3], as well as conditions such as Alzheimer's disease [4, 5], schizophrenia [6, 7], and multiple sclerosis [8, 9].

There are two main approaches for analyzing cortical thickness data: the region-level analysis and the whole-brain analysis (vertex-level). The region-level analysis first uses a brain parcellation atlas (e.g., Desikan-Killiany atlas) and obtains averaged cortical thickness data for each region of interest (ROI) [10, 11]. It offers the simplicity of achieving dimension reduction and alleviating multiple comparison problems better than whole-brain univariate analysis. However, it is limited to predefined regions, so it is unable to localize 'signal clusters' (patterns associated with certain conditions) that might emerge in smaller areas within ROIs or that span across multiple ROIs. Another approach is whole-brain analysis to analyze cortical thickness from all vertices throughout the brain [1, 12, 13], which offers better localization of signals at the expense of an increased burden in multiple comparisons [14]. On a positive note, recent studies on leveraging spatial dependencies inherent in vertex-level cortical thickness data have led to high power and effectively controlled the false positives [15, 16, 17]. Moreover, spatial covariance modelling in neuroimaging has shown evidence for promising performance in other neuroimaging modalities when integrated with cluster enhancement [18, 19].

Large-scale neuroimaging studies often use multi-site, multi-scanner protocols to recruit study participants quickly and in large numbers. However, a major challenge of combining neuroimaging studies across sites/scanners is *inter-scanner biases* that are introduced due to several technical variabilities in these studies, including disparities in scanner manufacturers, variations in scanner parameters, and heterogeneities in acquisition protocols [20, 21, 22]. As with other imaging modalities, these inter-scanner biases have been shown to be present in vertex-level cortical thickness data [21], which motivates a need for addressing inter-scanner biases and providing high-quality cortical thickness data for downstream whole-brain analysis.

Several statistical harmonization methods have been developed to identify and parameterize the source of inter-scanner biases and mitigate them by reconstructing new homogenized data for downstream analysis. One prominent approach, ComBat [23], first proposed in genomics, has been adapted for the removal of inter-scanner biases across various neuroimaging data modalities, including DTI mean diffusivity and fractional anisotropy [24], region-level cortical thickness [25], and functional connectivity [26]. ComBat characterizes scanner effects into an additive (mean)

and a multiplicative (variance) scanner effect for each imaging feature. Moreover, ComBat has been extended to harmonize imaging data collected in a longitudinal manner [27] and to harmonize MRI scans at the voxel level by incorporating the superpixel technique [28]. Recent harmonization methods (CovBat and RELIEF) have been shifted to expand the scope of statistical harmonization to address heterogeneous covariances, going beyond the mean-variance specifications in ComBat [29, 30]. However, their existing applications have primarily centred around regional-level neuroimaging data, and there is limited empirical evidence to assess their efficacy for vertex-level data. Also, the low-rank decomposition used by both CovBat and RELIEF might be sub-optimal and lose efficiency in vertex-level cortical thickness data, which exhibits significant spatial autocorrelation, which is because principal components loose rich local spatial information. Furthermore, CovBat and RELIEF do not preserve the spatial smoothness of the harmonized data, which would raise a critical issue when they are used in downstream analysis with spatial covariance modelling. This motivates a need for a new method that homogenizes spatial covariances of cortical thickness data in multi-site/-scanner studies and preserves the smoothness of the harmonized data.

To address these challenges, we propose a novel harmonization method called Spatial Autocorrelation Normalization (SAN) to identify and parameterize the sources of inter-scanner biases in vertex-level cortical thickness data and reconstruct homogenized and spatially smooth data. A central challenge of SAN lies in modelling scanner-specific covariances, differentiating them into heterogeneous non-spatial variations and spatial variations, all the while preserving the underlying homogeneous autocorrelation structures across scanners. In SAN, we use the spatial Gaussian process to leverage pairwise dependence in the characterization of covariance heterogeneity, which effectively addresses the local patterns of covariance heterogeneity and preserves spatial smoothness in the data. We propose a simple two-stage approach to model and estimate scanner-specific parameters for the heterogeneous means and covariances accordingly. Our method of moments (MoM) estimators provide a scalable and computationally efficient procedure for estimating these heterogeneous covariances. We apply our method to the Social Processes Initiative in the Neurobiology of the Schizophrenia(s) (SPINS) study, a multi-site, multi-scanner neuroimaging study including participants with schizophrenia spectrum disorders and healthy controls, to validate SAN, then construct a data-driven simulation to compare SAN to other harmonization methods.

2. Methods

2.1. Notations and model specifications

2.1.1. Characterization of heterogeneous means and variances

We let y_{ijv} be an imaging feature measured at vertex v ($v = 1, \ldots V$) of a hemisphere of the brain from subject j ($j = 1, \ldots, n_i$) in scanner i ($i = 1, \ldots, M$). Let $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijq})^\top$ be the

q-dimensional covariate vector for subject j in scanner i (e.g., age and sex). We model y_{ijv} by

$$\text{(Stage 1)} \quad y_{ijv} = \alpha_v + \mathbf{x}_{ij}^\top \boldsymbol{\beta}_v + \theta_{iv} + s_{iv}\epsilon_{ijv}, \tag{1}$$

where α_v is the intercept, $\boldsymbol{\beta}_v$ is the regression coefficient vector, θ_{iv} is the scanner-specific intercept for scanner i, s_{iv}^2 is the scanner-specific variance for scanner i, and ϵ_{ijv} is the noise term with the unit variance. Note that, at the vertex level, the Stage 1 model is equivalent to ComBat's specification of batch effects [23, 25] that models heterogeneous means and variances across scanners.

2.1.2. Characterization of heterogeneous spatial covariances

To account for spatial dependence of cortical thickness, we model $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijV})^\top$ using the spatial Gaussian process (GP), that is, $\boldsymbol{\epsilon}_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{MVN}(\mathbf{0}_V, \boldsymbol{\Sigma}_i)$ for $j = 1, \ldots, n_i$. Specifically, we decompose $\boldsymbol{\epsilon}_{ij}$ into three additive random effects as

$$\text{(Stage 2)} \quad \boldsymbol{\epsilon}_{ij} = \boldsymbol{\gamma}_{ij}^S + \boldsymbol{\gamma}_{ij}^E + \boldsymbol{\delta}_{ij}, \tag{2}$$

with an assumption that $\boldsymbol{\gamma}_{ij}^S, \boldsymbol{\gamma}_{ij}^E, \boldsymbol{\delta}_{ij}$ are independent to each other.

- $\boldsymbol{\gamma}_{ij}^S$ and $\boldsymbol{\gamma}_{ij}^E$ are spatial random effects that are modeled by $\boldsymbol{\gamma}_{ij}^S \sim \mathcal{MVN}(\mathbf{0}_V, \sigma_{S,i}^2 \cdot \boldsymbol{\Phi}_S(\phi_S))$ and $\boldsymbol{\gamma}_{ij}^E \sim \mathcal{MVN}(\mathbf{0}_V, \sigma_{E,i}^2 \cdot \boldsymbol{\Phi}_E(\phi_E))$. Here, we assume that the covariances of $\boldsymbol{\gamma}_{ij}^S$ and $\boldsymbol{\gamma}_{ij}^E$ are characterized by (i) heterogeneous spatial variances which represented by scanner-specific parameters $\sigma_{S,i}^2$ and $\sigma_{E,i}^2$ and (ii) homogeneous underlying autocorrelations across scanners which represented by the spatial autocorrelation functions (SACFs). For vertices v and v^*,

$$\begin{aligned} \text{(Squared exponential SACF)} \quad & \boldsymbol{\Phi}_S(\phi_S)[v, v^*] = \exp(-\phi_S \cdot d_{v,v^*}^2), \\ \text{(Exponential SACF)} \quad & \boldsymbol{\Phi}_E(\phi_E)[v, v^*] = \exp(-\phi_E \cdot d_{v,v^*}), \end{aligned}$$

  where d_{v,v^*} is the geodesic distance between vertices v and v^*, and the scanner-invariant parameters ϕ_S and ϕ_E determine how fast the spatial autocorrelation decreases with distance. The exponential SACF falls off rapidly with small distances but then tails off much slower than the squared exponential SACF as distance increases. Combining these two forms provides the flexibility to simultaneously account for spatial correlation at shorter distances through the squared exponential SACF and capture the heavy-tailed nature of spatial dependence within the brain through the exponential SACF. The integration of squared exponential and exponential SACFs has shown its utility when modeling spatial autocorrelations in fMRI data [31].

- $\boldsymbol{\delta}_{ij}$ is the non-spatial effect modeled by $\boldsymbol{\delta}_{ij} \sim \mathcal{MVN}(\mathbf{0}_V, \tau_i^2 \mathbf{I}_V)$. Its covariance structure $\tau_i^2 \mathbf{I}_V$ includes scanner-specific parameter τ_i^2 represents heterogeneous non-spatial variances across scanners.

Altogether, the marginal covariance of $\boldsymbol{\epsilon}_{ij}$ is $\boldsymbol{\Sigma}_i = \sigma^2_{S,i}\boldsymbol{\Phi}_S(\phi_S) + \sigma^2_{E,i}\boldsymbol{\Phi}_E(\phi_E) + \tau_i^2\mathbf{I}_V$.

2.2. Spatial autocorrelation normalization (SAN)

2.2.1. Stage 1

The Stage 1 model requires estimating α_v, $\boldsymbol{\beta}_v$, θ_{iv} and s^2_{iv}, with an important data-specific consideration that the additive and multiplicative batch effects (θ_{iv} and s^2_{iv}) would be smooth over space. Considering such smoothness in the estimation step would reduce the variance of the parameter estimates, which is expected to improve the performance of the Stage 2 of SAN especially when the number of subjects is small. To achieve this, SAN pools information within a series of prespecified local neighbors [32]. Specifically, we define local neighbors $\mathcal{N}_r(v)$ for each vertex as the set of vertices whose geodesic distances from vertex v are less than or equal to r.

- **Collapsing**: The first method is to apply average smoothing locally on $\mathcal{N}_r(v)$. After constructing vertex-wise linear regressions between y_{ijv} and $\mathbf{x}_{ij}$ to obtain $\hat{\alpha}_v$, $\hat{\boldsymbol{\beta}}_v$ and residuals $\hat{e}_{ijv} = y_{ijv} - \hat{\alpha}_v - \mathbf{x}_{ij}^\top\hat{\boldsymbol{\beta}}_v$, we iteratively obtain $\hat{\theta}_{iv}$, $\hat{s}^2_{iv}$, and $\hat{\epsilon}_{ijv}$ by

$$\hat{\theta}_{iv} = \frac{1}{n_i|\mathcal{N}_r(v)|}\sum_{v'\in\mathcal{N}_r(v)}\sum_{j=1}^{n_i}\hat{e}_{ijv^*} \quad \text{and} \quad \hat{s}^2_{iv} = \left(\frac{1}{n_i|\mathcal{N}_r(v)| - M - q}\sum_{v^*\in|\mathcal{N}_r(v)|}\sum_{j=1}^{n_i}(\hat{e}_{ijv^*} - \hat{\theta}_{iv})^2\right)$$

  and obtain $\hat{\epsilon}_{ijv} = \frac{\hat{e}_{ijv}-\hat{\theta}_{iv}}{\hat{s}_{iv}}\sqrt{\frac{n_i|\mathcal{N}_r(v)|}{n_i|\mathcal{N}_r(v)|-M-q}}$.

- **ComBat**: We apply ComBat [23, 25] to each $\mathcal{N}_r(v)$. In ComBat, we obtain $\hat{\alpha}_v$, $\hat{\boldsymbol{\beta}}_v$ and residuals $\hat{e}_{ijv}$ using the same way as the Collapsing model. However, in contrast to the collapsing approach, which assumes that local neighbors share identical values of θ_{iv} and s^2_{iv}, ComBat imposes normal-inverse-gamma priors to θ_{iv} and s^2_{iv} and estimates them through the empirical Bayes approach, which would shrink these estimates towards the overall means and variances of $\mathcal{N}_r(v)$.

The difference between local collapsing and local ComBat is whether the fixed effect or random effect is used to achieve smooth estimates of θ_{iv} and s^2_{iv}. In both approaches, the performance of the proposed method depends on the selection of the radius r. When $r = 0$, we estimate θ_{iv} and s^2_{iv} separately for each vertex without borrowing any information from local neighbors. When $r = \infty$, we estimate θ_{iv} and s^2_{iv} by using all vertices in the data, which would result in 'too smooth' estimates shrunk towards the brain-level means and variances. An appropriate level of r needs to be chosen by considering the bias-variance tradeoff [32]. In this paper, we choose ComBat model with $r = 5$mm as a default to borrow information across neighbors without inducing too much bias.

2.2.2. Stage 2

We generalize the covariance regression proposed by Zou et al. [33] to estimate $(\phi_S, \phi_E, \sigma^2_{S,i}, \sigma^2_{E,i}, \tau^2_i)$. The covariance regression approach uses the method of moments (MoM) estimators, which provide a highly computationally efficient and consistent estimator of the variance component parameters, with some loss in efficiency compared to the likelihood-based methods. We note that the likelihood-based methods would be intractable as the number of vertices gets larger or the number of scanners/sites increases [33]. Previous work from Park and Fiecas [18] has also shown empirically that the MoM estimators yielded nearly unbiased estimates of parameters. In SAN, the parameters are estimated by minimizing the following objective function:

$$\{\hat{\phi}_S, \hat{\phi}_E, \hat{\sigma}^2_{S,i}, \hat{\sigma}^2_{E,i}, \hat{\tau}^2_i\} = \underset{\{\phi_S, \phi_E, \sigma^2_{S,i}, \sigma^2_{E,i}, \tau^2_i\}}{\arg\min} \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \|\hat{\boldsymbol{\epsilon}}_{ij}\hat{\boldsymbol{\epsilon}}_{ij}^\top - \sigma^2_{S,i}\boldsymbol{\Phi}_S(\phi_S) - \sigma^2_{E,i}\boldsymbol{\Phi}_E(\phi_E) - \tau^2_i \mathbf{I}_V\|_F^2 \right\}, \tag{3}$$

where $||\cdot||_F^2$ is the squared Frobenius norm of a matrix. For optimization, we first note that, when ϕ_S and ϕ_E are given (so that $\boldsymbol{\Phi}_G \equiv \boldsymbol{\Phi}_G(\phi_G)$ and $\boldsymbol{\Phi}_E \equiv \boldsymbol{\Phi}_E(\phi_E)$), a closed-form solution for other parameters minimizing the objective (3) is provided by

$$\begin{pmatrix} \hat{\sigma}^2_{S,1} \\ \hat{\sigma}^2_{E,1} \\ \hat{\tau}^2_1 \\ \vdots \\ \hat{\sigma}^2_{S,M} \\ \hat{\sigma}^2_{E,M} \\ \hat{\tau}^2_M \end{pmatrix} = \begin{pmatrix} n_1\mathrm{tr}(\boldsymbol{\Phi}_S\boldsymbol{\Phi}_S) & n_1\mathrm{tr}(\boldsymbol{\Phi}_S\boldsymbol{\Phi}_E) & n_1V & \dots & 0 & 0 & 0 \\ n_1\mathrm{tr}(\boldsymbol{\Phi}_S\boldsymbol{\Phi}_E) & n_1\mathrm{tr}(\boldsymbol{\Phi}_E\boldsymbol{\Phi}_E) & n_1V & \dots & 0 & 0 & 0 \\ n_1V & n_1V & n_1V & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & n_M\mathrm{tr}(\boldsymbol{\Phi}_S\boldsymbol{\Phi}_S) & n_M\mathrm{tr}(\boldsymbol{\Phi}_S\boldsymbol{\Phi}_E) & n_MV \\ 0 & 0 & 0 & \dots & n_M\mathrm{tr}(\boldsymbol{\Phi}_S\boldsymbol{\Phi}_E) & n_M\mathrm{tr}(\boldsymbol{\Phi}_E\boldsymbol{\Phi}_E) & n_MV \\ 0 & 0 & 0 & \dots & n_MV & n_MV & n_MV \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{n_1} \boldsymbol{\epsilon}_{1j}^\top\boldsymbol{\Phi}_S\boldsymbol{\epsilon}_{1j} \\ \sum_{j=1}^{n_1} \boldsymbol{\epsilon}_{1j}^\top\boldsymbol{\Phi}_E\boldsymbol{\epsilon}_{1j} \\ \sum_{j=1}^{n_1} \boldsymbol{\epsilon}_{1j}^\top\boldsymbol{\epsilon}_{1j} \\ \vdots \\ \sum_{j=1}^{n_M} \boldsymbol{\epsilon}_{Mj}^\top\boldsymbol{\Phi}_S\boldsymbol{\epsilon}_{Mj} \\ \sum_{j=1}^{n_M} \boldsymbol{\epsilon}_{Mj}^\top\boldsymbol{\Phi}_E\boldsymbol{\epsilon}_{Mj} \\ \sum_{M=1}^{n_M} \boldsymbol{\epsilon}_{Mj}^\top\boldsymbol{\epsilon}_{Mj} \end{pmatrix}.$$

Subsequently, when we plug in $(\hat{\sigma}^2_{S,i}, \hat{\sigma}^2_{E,i}, \hat{\tau}^2_i)$ to the closed-form solution in Equation (3), it simplifies the optimization process, reducing it to finding the minimum of the updated objective function with respect to (ϕ_S, ϕ_E). This step reduces the number of parameters for the optimization from $2 + 3M$ to 2, significantly enhancing the computational efficiency regardless of the number of scanners (M) used in a study. In our implementation in `R`, the Nelder-Mead method [34] was used to solve this nonlinear optimization problem.

Under the GP assumption, the conditional expectations of $\boldsymbol{\gamma}^S_{ij}$, $\boldsymbol{\gamma}^E_{ij}$ and $\boldsymbol{\delta}_{ij}$ are given by

$$\begin{aligned} \hat{\boldsymbol{\gamma}}^S_{ij} &= \widehat{\mathrm{E}}[\boldsymbol{\gamma}^S_{ij}|\hat{\boldsymbol{\epsilon}}_{ij}] = \hat{\sigma}^2_{S,i}\boldsymbol{\Phi}_S(\hat{\phi}_S)\widehat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\epsilon}}_{ij}, \\ \hat{\boldsymbol{\gamma}}^E_{ij} &= \widehat{\mathrm{E}}[\boldsymbol{\gamma}^E_{ij}|\hat{\boldsymbol{\epsilon}}_{ij}] = \hat{\sigma}^2_{E,i}\boldsymbol{\Phi}_E(\hat{\phi}_E)\widehat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\epsilon}}_{ij}, \\ \hat{\boldsymbol{\delta}}_{ij} &= \widehat{\mathrm{E}}[\boldsymbol{\delta}_{ij}|\hat{\boldsymbol{\epsilon}}_{ij}] = \hat{\tau}^2_i\widehat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\epsilon}}_{ij}. \end{aligned}$$

2.2.3. Construction of harmonized data

With all decompositions and estimations made in both stages, we remove scanner-specific means and normalize scanner-specific covariances to construct the harmonized data. Since the sources of covariance heterogeneity are characterized by scanner-specific parameters $(\sigma^2_{S,i}, \sigma^2_{E,i}, \tau^2_i)$, we normalize $\hat{\boldsymbol{\gamma}}^S_{ij}$, $\hat{\boldsymbol{\gamma}}^E_{ij}$ and $\hat{\boldsymbol{\delta}}_{ij}$ by

$$\boldsymbol{\gamma}^{S(n)}_{ij} = \frac{\sigma^{(n)}_S}{\hat{\sigma}_{S,i}}\hat{\boldsymbol{\gamma}}^S_{ij}, \quad \boldsymbol{\gamma}^{E(n)}_{ij} = \frac{\sigma^{(n)}_E}{\hat{\sigma}_{E,i}}\hat{\boldsymbol{\gamma}}^E_{ij}, \quad \boldsymbol{\delta}^{(n)}_{ij} = \frac{\tau^{(n)}}{\hat{\tau}_i}\hat{\boldsymbol{\delta}}_{ij},$$

where $\sigma^{(n)}_S = \sqrt{(\sum_i n_i \cdot \hat{\sigma}^2_{S,i})/n}$, $\sigma^{(n)}_E = \sqrt{(\sum_i n_i \cdot \hat{\sigma}^2_{E,i})/n}$ and $\tau^{(n)} = \sqrt{(\sum_i n_i \cdot \hat{\tau}^2_i)/n}$. Therefore, the final normalized data is given by

$$y^{(n)}_{ijv} = \hat{\alpha}_v + \mathbf{x}^\top_{ij}\hat{\boldsymbol{\beta}}_v + s^{(n)}_v \times (\gamma^{S(n)}_{ijv} + \gamma^{E(n)}_{ijv} + \delta^{(n)}_{ijv}),$$

where $s^{(n)}_v = \sqrt{(\sum_i n_i \cdot \hat{s}^2_{iv})/n}$.

2.3. Integrating SAN with other harmonization methods

Stage 2 of SAN provides a model-based framework for characterizing scanner-specific covariances into heterogeneous spatial variations (modeled by $\boldsymbol{\gamma}^S_{ij}$ and $\boldsymbol{\gamma}^E_{ij}$) and non-spatial variations (modeled by $\boldsymbol{\delta}_{ij}$). While the primary focus of SAN lies in modelling and normalizing scanner-specific spatial autocorrelations, it is worth noting that spatial heterogeneity might not be the only source of the scanner effects. In such a case, SAN's formulation could suffer from oversimplification, failing to address the full complexity associated with heterogeneous non-spatial variations. Therefore, we considered applying covariance harmonization methods to $\hat{\boldsymbol{\delta}}_{ij}$ from Stage 2 of SAN to capture potential remaining scanner effects. In this paper, we consider CovBat [29] ("SAN+CovBat") and RELIEF [30] ("SAN+RELIEF") as these are methods primarily developed to harmonize covariances.

3. Data Analysis

3.1. Data preparation and preprocessing

We used cortical thickness data from the Social Processes Initiative in the Neurobiology of the Schizophrenia(s) (SPINS) study to evaluate SAN's performance. The SPINS study includes multimodal neuroimaging data in individuals diagnosed with schizophrenia spectrum disorders (SSDs) and control participants. Participants aged 18-59 were recruited for SPINS between 2014-2020. 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. See Viviano et al. [35] and Oliver et al. [36] for details.

Scans were obtained from three different imaging sites: Centre for Addiction and Mental Health (CAMH), Maryland Psychiatric Research Center (MPRC), and Zucker Hillside Hospital (ZHH). General Electric (GE) 3T MRI scanners were used at CAMH and ZHH (750w Discovery and Signa, respectively), while MPRC used the Siemens Tim Trio 3T (ST). However, in the third year of the study, all sites switched to the Siemens Prisma (SP) scanner. Given the limited number of samples from the Siemens Tim Trio 3T (60 samples), we excluded ST data and focused our analysis on the two scanner types: GE and SP. Our final dataset comprised 357 subjects across these two scanner types: 164 were imaged on the GE (66 females, 104 patients, aged 18-55) and 193 on the SP (79 females, 105 patients, aged 18-55) after visual quality control.

MRI data preprocessing was performed using fMRIPrep 1.5.8 [37], based on Nipype 1.4.1 [38]. T1-weighted images were corrected for intensity non-uniformity and skull-stripped using ANTs [39]. Cortical surfaces were reconstructed using FreeSurfer 6.0.1 [40]. Cortical thickness data were resampled to fsaverage5 space, including gaussian smoothing with FWHM of 0, 5, and 10 mm, leaving 9,354 cortical vertices for the left hemisphere and 9,361 for the right hemisphere after excluding vertices on the medial wall. Upon visual inspection of the preprocessed T1 images, we excluded 29 images that did not pass quality control or passed with small issues, including minor bad skull stripping, MNI warping, and tiny under-inclusive Freesurfer masking. The pairwise geodesic distance was computed from the pial surface to reflect surface geometry.

3.2. Smoothing increases heterogeneity in covariances across scanners

In terms of cortical thickness data preprocessing pipelines, there is currently no consensus on whether smoothing should be done before or after harmonization However, our study provides empirical evidence that spatial smoothing can exacerbate the complexity of covariance heterogeneity among different scanners. To quantify the inter-scanner variabilities in covariances, we introduce a measure called Covar$F$ statistic that extends the idea of F statistic. This measure is defined as the ratio of inter-scanner variabilities and within-scanner variabilities in empirical covariances given by

$$\text{Covar}F(\mathcal{N}) = \frac{\sum_{(v,v^*)\in\mathcal{N}}(\sum_{i=1}^{M} n_i(\hat{\sigma}_i(v,v^*) - \hat{\sigma}(v,v^*))^2)/(M-1)}{\sum_{(v,v^*)\in\mathcal{N}}(\sum_{i=1}^{M}\sum_{j=1}^{n_i}(\hat{\sigma}_{ij}(v,v^*) - \hat{\sigma}_i(v,v^*))^2)/(n-M)}, \tag{4}$$

where $\hat{\sigma}_{ij}(v,v^*) = (y_{ijv} - \hat{\alpha}_v - \mathbf{x}_{ij}^\top\hat{\boldsymbol{\beta}}_v - \hat{\theta}_{iv})(y_{ijv^*} - \hat{\alpha}_{v^*} - \mathbf{x}_{ij}^\top\hat{\boldsymbol{\beta}}_{v^*} - \hat{\theta}_{iv^*})$. Here, $\hat{\sigma}_i(v,v^*) = \frac{1}{n_i}\sum_j \hat{\sigma}_{ij}(v,v^*)$ represents the covariance of cortical thickness between vertex v and v^* for all subjects within scanner i, and $\hat{\sigma}(v,v^*) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\hat{\sigma}_{ij}(v,v^*)$ represents the covariance between vertex v and v^* for pooled cortical thickness data across both subjects and scanners. $\mathcal{N}$ is the set of vertex pairs that meet certain conditions, and we will provide several variants of $\mathcal{N}$ in the following sections. For the current context, we define $\mathcal{N} = \mathcal{N}_r(v)$ with $r = 5$mm, denoting the local neighbors for each vertex v. Then we map these Covar$F$ statistics onto the inflated brain surface. A severe inter-scanner bias in covariances is reflected in the larger Covar$F$ statistic.

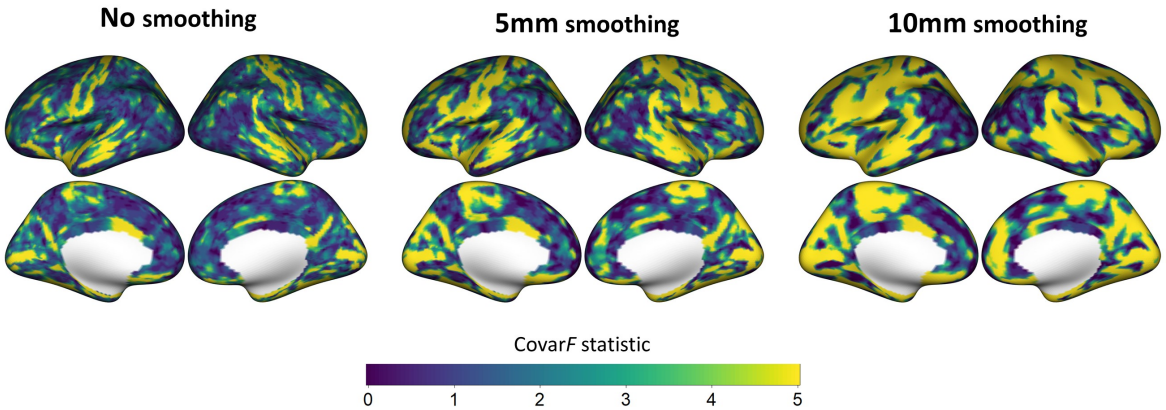

Figure 1: Covar$F$ statistic brain maps obtained from unsmoothed and smoothed data before applying any harmonization methods. A larger Covar$F$ statistic at a vertex implies higher covariance heterogeneity in local neighbors surrounding the vertex.

In our experiments, we use three sets of cortical thickness data: unsmoothed data, 5mm smoothed data and 10mm smoothed data. From Figure 1, we first observe from results of the unsmoothed data that covariance heterogeneity is localized and prominent in regions including, but not limited to, pericalcarine, caudal anterior cingulate, paracentral, precentral, postcentral, superior temporal, midtemporal, and insula and entorhinal cortices. We also see that smoothing not only intensifies inter-scanner covariance biases but also spreads these biases to a larger spatial extent compared to unsmoothed data. This effect is further magnified when a larger smoothing kernel is used. We illustrate in a simplified setting with mathematical proof in Appendix A to show how spatial smoothing can amplify the covariance heterogeneity between different scanners. With all these results, we recommend performing data harmonization before spatial smoothing. Therefore, SAN's performance is evaluated using the unsmoothed data throughout this paper, with sensitivity analysis regarding the effect of smoothing provided in Section 3.6.

3.3. Brain-level analysis

3.3.1. Covariogram analysis

Covariograms quantify spatial dependence and variability between locations. Specifically, for subject j from scanner i, the covariogram is defined as:

$$C_{ij}(d) = \frac{1}{|\mathcal{N}(d)|} \sum_{(v,v^*)\in\mathcal{N}(d)} \hat{\sigma}_{ij}(v, v^*), \tag{5}$$

where $\hat{\sigma}_{ij}(v, v^*)$ is as provided in Section 3.2. Here, $\mathcal{N}(d)$ is the set of vertex pairs such that their geodesic distance is between $d - \eta$ and $d + \eta$ for a small bandwidth $\eta > 0$. $|\mathcal{N}(d)|$ is the number of elements in this set. Similarly, we also applied Equation (5) to the harmonized data (denoted by $C_{ij^{(n)}}$), by replacing y_{ijv} with $y_{ijv}^{(n)}$ in the same formula and subsequently deriving $C_{ij}^{(n)}(d)$ by using $\hat{\sigma}_{ij}^{(n)}$. To assess spatial variations in the pooled original data, we obtain the averaged empirical covariogram across scanners, known as $C(d)$, by $C(d) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{n_i} C_{ij}(d)$. We also obtain scanner-averaged empirical covariograms for harmonized data analogously by $C_i^{(n)}(d) = \frac{1}{n_i}\sum_j C_{ij}^{(n)}(d)$.

To visualize whether spatial variations are homogeneous across scanners after harmonization, we compute covariogram ratios, $C_i^{(n)}(d)/C(d)$, between $C_i^{(n)}(d)$ using different harmonization methods and $C(d)$ for pooled original data. Because $C(d)$ considers averaged covariogram across scanners and subjects (after regressing out scanner-specific means), a harmonization method that homogenizes spatial covariances better would result in the scanner-averaged covariogram shrunk $C_i^{(n)}(d)$ towards $C(d)$, leading to a ratio closer to 1. In Figure 2(a), when considering the covariogram ratios for both GE and SP within the range of 0mm to 40mm, SAN consistently outperforms the other methods across most distances. While CovBat effectively aligns the SP-averaged covariogram closer to the pooled covariogram, it introduces more deviations in the covariogram ratios for GE when distances exceed 20mm. AdjRes and ComBat exhibit similar covariogram ratios, consistently larger than those of SAN. Notably, RELIEF displays distinct behavior compared to the other methods; it demonstrates the smallest ratios for SP but the largest ratios for GE, even surpassing AdjRes in the latter case. This observation suggests that RELIEF's performance exhibits an imbalance across scanners, potentially arising from an excessive elimination of spatial variation specific to GE.

3.3.2. CovarF statistic

To quantify how much harmonization methods reduce the inter-scanner variabilities in covariances, we also use Covar$F$ statistics in Equation (4) with some modifications; specifically, we replace $\hat{\sigma}_{ij}(v, v^*)$ and $\hat{\sigma}_i(v, v^*)$ with $\hat{\sigma}_{ij}^{(n)}(v, v^*)$ and $\hat{\sigma}_i^{(n)}(v, v^*)$ for each harmonized data, and we use $\mathcal{N} = \mathcal{N}(d)$ as provided in Section 3.3.1. This allows us to evaluate the average degree of covariance heterogeneity in the brain depending on distances, and effectiveness in mitigating inter-scanner biases in covariances is reflected in smaller Covar$F$ statistics. The unitless property of this Covar$F$ statistic allows for easy interpretation and comparison of different harmonization methods.

Figure 2(b) presents Covar$F$ statistics across distances from 0mm to 40mm. As shown in Figure 2(b), noticeable performance differences among harmonization methods are observed within the 0-20mm range. Specifically, SAN, SAN+RELIEF, SAN+CovBat, and CovBat exhibit markedly smaller Covar$F$ statistics compared to other methods. Within this range, our methods outperform CovBat within the 10mm interval. However, for distances ranging from 10mm to 20mm, CovBat shows slightly smaller Covar$F$ statistics. Beyond the 20mm threshold, all methods demon-

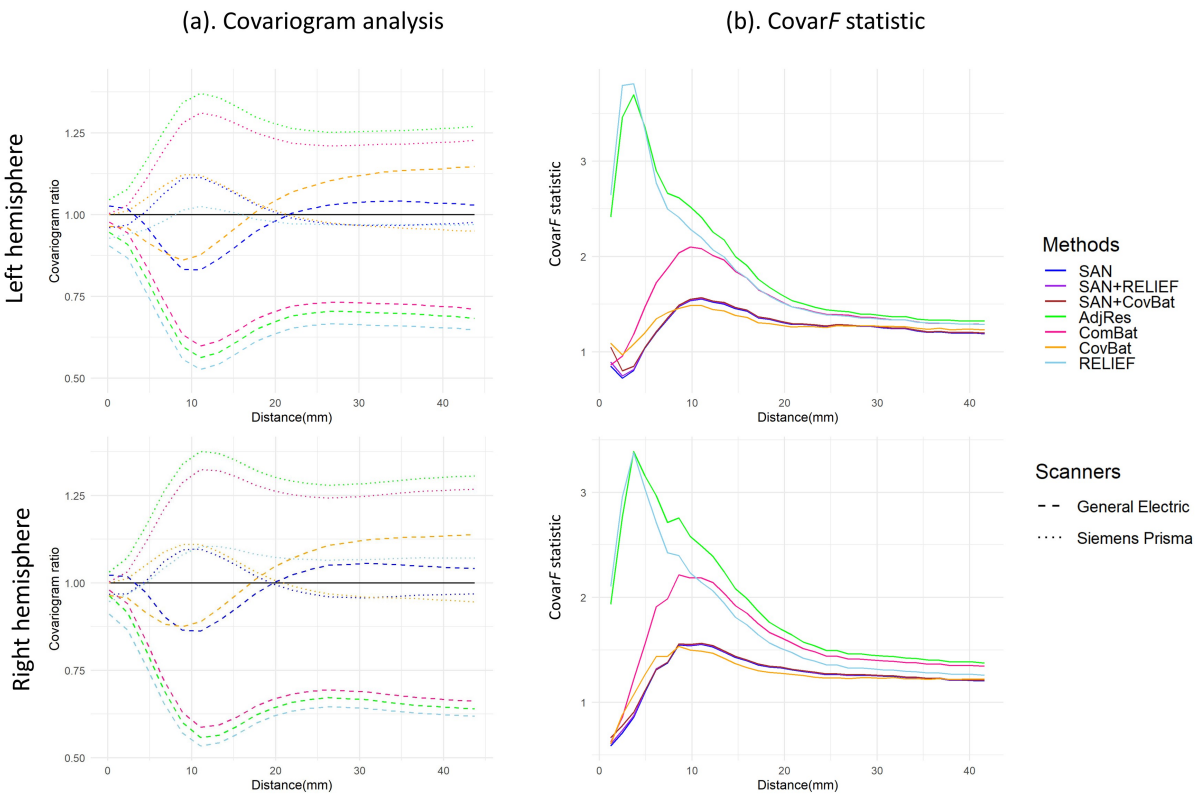

Figure 2: Summary of covariogram ratios and Covar$F$ statistics obtained from different harmonization methods. The left panel shows covariogram ratios, with the solid horizontal line representing a ratio of 1. The results SAN+RELIEF and SAN+CovBat are not shown as they exhibited the same covariogram ratios as SAN. The right panel shows Covar$F$ statistics.

strate similar performance, with Covar$F$ statistics decreasing and then stabilizing. It could be because (spatial) heterogeneity in covariances would be marginal as the distance increases. Still, SAN, SAN+RELIEF and SAN+CovBat consistently produce slightly smaller Covar$F$ statistics, supporting the empirical performance of SAN.

3.3.3. CovarF statistic brain maps

To clearly visualize the decrease in covariance heterogeneity between scanners both globally and locally on the brain surface after harmonization, we compute Covar$F$ statistics for each harmonization method as done in Section 3.2. Figure 3 shows the brain maps of Covar$F$ statistics. Overall, these maps generally show similar patterns and are also similar to Covar$F$ map in the unsmoothed data from Figure 1, with pronounced Covar$F$ statistics concentrated around specific spatial clusters. Notably, SAN, SAN+RELIEF and SAN+CovBat show great improvement in reducing Covar$F$ statistics compared to other methods. Covariance heterogeneity between scanners

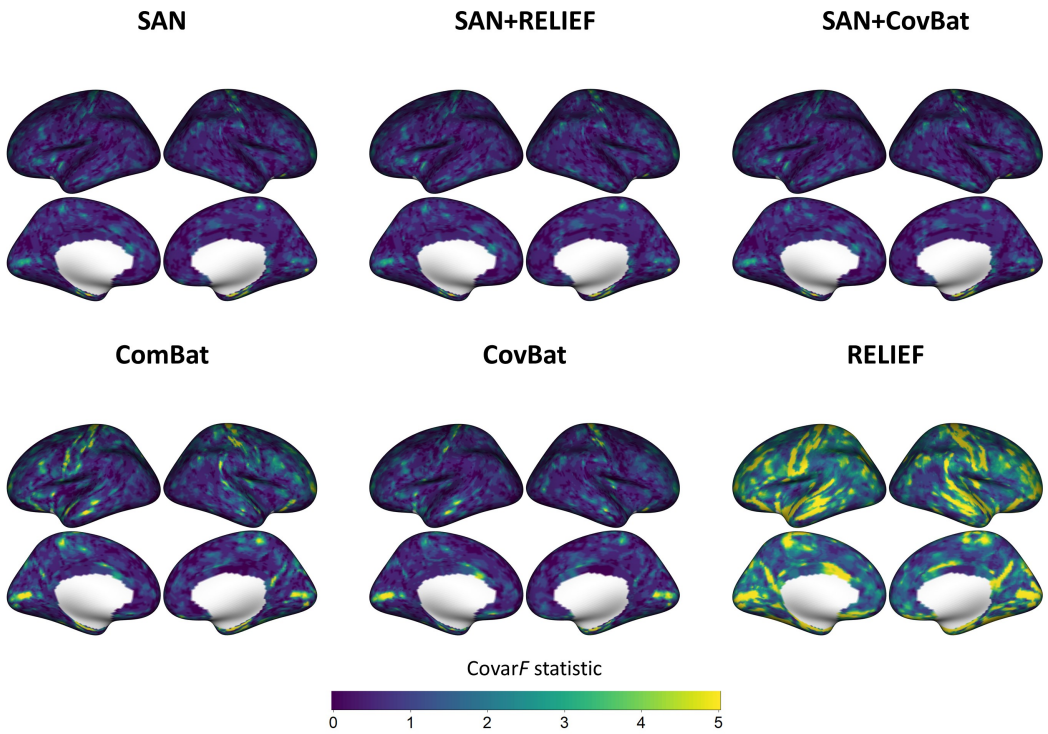

Figure 3: Covar$F$ statistic brain maps obtained from different harmonization methods.

remains prominent in RELIEF brain maps. ComBat and CovBat perform slightly better, though some minor spatial clusters still present large Covar$F$ statistics.

3.4. Within-ROI analysis

The purpose of this section is to investigate how brain-wise spatial harmonization affects locally within a predefined region. We use Desikan-Killiany Atlas [41], which includes 34 regions of interest (ROIs) in each hemisphere (after excluding corpus callosum). With each harmonized dataset, we segment the cortical thickness values into 34 subsets for the left hemisphere according to this atlas. To evaluate and compare the performance of harmonization methods in reducing the inter-scanner variabilities in covariances across different ROIs, we also use Covar$F$ statistics but averaging across each region at a time (e.g., $\mathcal{N} = \mathcal{N}(R)$, denoting the set of vertex pairs within region R).

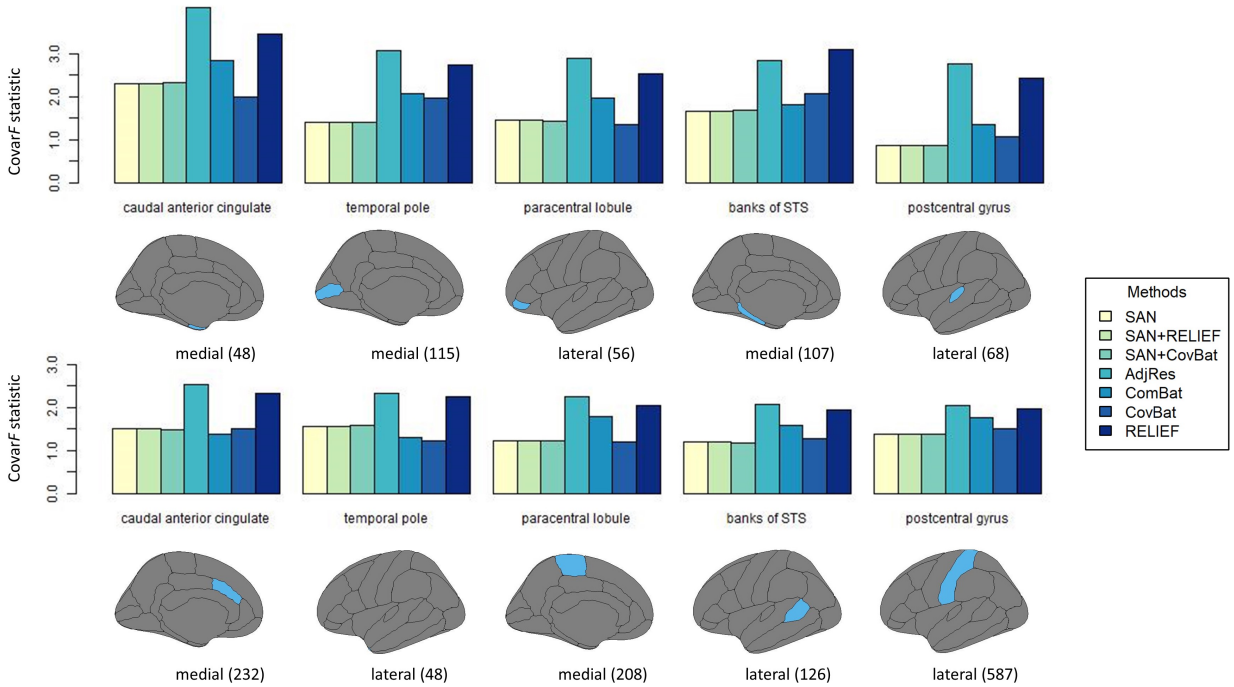

Figure 4: Covar$F$ statistics obtained from different harmonization methods for the top 10 regions with the highest Covar$F$ statistics in the AdjRes harmonized data. Below the corresponding bar plots, diagrams of the ROI complete with vertex counts are displayed.

Figure 4 shows the barplots of Covar$F$ statistics for the top 10 regions with the highest Covar$F$ statistics from the original data (which provides the same Covar$F$ statistic as AdjRes). Overall, these regions are relatively smaller in size, having 142 vertices on average, which is less than the average of 265 vertices for all regions. This observation is also consistent with our findings in section 3.3, where the inter-scanner heterogeneity in covariances is most pronounced at shorter distances. Also, among these regions, SAN, SAN+RELIEF and SAN+CovBat show the lowest Covar$F$ statistics for 6 regions. CovBat showed the lowest Covar$F$ statistics for 3 regions, which partially supports that there might be some degrees of spatial nonstationarity in cortical thickness data such that SAN (which assumes spatial stationarity) would perform worse than CovBat in some localized areas. Lastly, the performance of ComBat was poor in harmonizing covariances within most ROIs. While ComBat shows the lowest Covar$F$ statistic for caudal anterior cingulate, the disparities between ComBat and SAN are marginal.

3.5. Between-ROI analysis

The purpose of this section is to investigate whether brain-wise harmonization before parcellation is effective in homogenizing covariances in region-level data. Using Desikan-Killiany atlas,

we derive ROI-level cortical thickness data by averaging cortical thickness values within each corresponding region. In this section, we assess two types of harmonized data: (i) the vertex-level harmonization followed by parcellation and (ii) partitioning the data to construct the average cortical thickness for each ROI and then harmonizing the ROI-level cortical thickness data. For (ii), we use ComBat, CovBat, and RELIEF (denoted as ROI-level ComBat, ROI-level CovBat and ROI-level RELIEF). Although vertex-level data may not be directly applicable to ROI-level analyses, the advantages observed in capturing and retaining the underlying shared spatial correlation across scanners can still be leveraged when working with ROI-level data by representing averaged measurements within regions.

In addition to the Covar$F$ statistic, we also use Quadratic Discriminant Analysis (QDA) to assess the predictive capability of each harmonized data in relation to scanners. It is important to note that a harmonization method that exhibits superior performance in mitigating scanner effects will yield inferior predictive performance. Similarly to Zhang et al. [30], we choose QDA because it relies solely on mean vectors and covariance matrices. Therefore, any variations in predictive performance can be directly attributed to the harmonization of scanner-specific means and covariances. Using leave-one-out cross-validation, we calculate the average accuracy as well as ROC curve's area under the curve (AUC) for each harmonized dataset after regressing out covariate effects.

The results for the Covar$F$ statistics, accuracy and AUC for scanner prediction are shown in Table 1. For Covar$F$ statistics, the ROI-level ComBat and ROI-level CovBat exhibit the most effective performance in minimizing inter-scanner variabilities. It can be attributed to the favorable utilization of ROI-level data in ROI-level ComBat/CovBat for harmonization. However, SAN, SAN+RELIEF, and SAN+CovBat show noticeably smaller Covar$F$ statistics than all other vertex-level harmonization methods. This suggests that our methods excel in recovering underlying homogeneous spatial correlations across scanners. For evaluating the predictive performance of scanners, we see that ROI-level RELIEF and vertex-level RELIEF achieve the lowest prediction accuracy and AUC values, which is consistent with the findings by Zhang et al. [30]. Considering that RELIEF performed worse in reducing the Covar$F$ statistics, this result suggests that an optimal harmonization method should be chosen carefully based on the purpose of the desired research (e.g., prediction vs. inference).

3.6. Sensitivity analysis of the impact of smoothing

As discussed in Section 3.2, researchers may perform spatial smoothing before harmonization. To understand the impact of smoothing on harmonization performance, we investigate the sensitivity of harmonization methods to varying levels of smoothing, specifically at 5mm and 10mm. We again use Covar$F$ statistics as defined in Section 3.2. Then we map their Covar$F$ statistics onto the inflated brain surface.

Measure	Covar$F$	Accuracy	AUC
SAN	**2.351**	0.608	0.601
SAN+RELIEF	2.352	0.608	0.601
SAN+CovBat	2.360	0.611	0.603
AdjRes	4.546	0.641	0.628
ComBat	3.648	0.605	0.593
CovBat	4.152	0.622	0.61
RELIEF	4.349	**0.529**	**0.517**
ROI-level ComBat	**1.746**	0.605	0.602
ROI-level CovBat	1.823	0.555	0.549
ROI-level RELIEF	3.827	**0.527**	**0.517**

Table 1: The summary of Covar$F$ statistic, the accuracy and AUC of predicting scanners. The smallest values for vertex-level/ROI-level methods are bolded respectively.

Figure 5 shows Covar$F$ statistic brain maps for both 5mm smoothed and 10mm smoothed harmonized data. Brain maps corresponding to the 5mm smoothing level are similar to those observed in unsmoothed harmonized data. SAN, SAN+RELIEF and SAN+CovBat outperform other harmonization methods in reducing inter-scanner variabilities. However, these perform poorly with the 10mm smoothing, as indicated by the appearance of large spatial clusters characterized by notably high Covar$F$ statistics on the maps. This limitation is likely due to the nonstationarity induced by spatial smoothing, which could lead to a severe violation of the stationary Gaussian process assumption made by SAN. Using a smoothing kernel size of 10mm may include multiple anatomically distinct regions, and it might introduce or exaggerate nonstationarity at broader scales. It is also worth noting that ComBat (which does not address covariance heterogeneity) seems to be the best model for harmonizing covariances when 10mm smoothing is applied. Therefore, we note that SAN is preferably applied to unsmoothed data or data with minimal smoothing levels (e.g., less than 5mm) to ensure optimal harmonization outcomes.

3.7. Data-driven simulations

The Covar$F$ statistics of SPINS cortical thickness data are based on a single realization, which is inadequate to draw conclusions about its precision and uncertainty. The underlying population distribution of SPINS cortical thickness data may reveal nonstationarity, making evaluations through simulations with stationary assumptions less suitable. To address these limitations, we perform 1000 bootstraps. Bootstrapping allows us to draw inferences about precision and uncertainty directly from the observed data. Additionally, it ensures that our analysis remains data-driven and reflects the characteristics of the actual dataset. Each bootstrap involves randomly selecting 75 individuals from GE and 75 individuals from SP. To streamline computations and enhance result presentation, we introduce some modifications in the calculation of Covar$F$ statistics. Given the large number of bootstrapping samples, we select distance ranges as: 0mm-5mm, 5mm-10mm, 10mm-15mm,

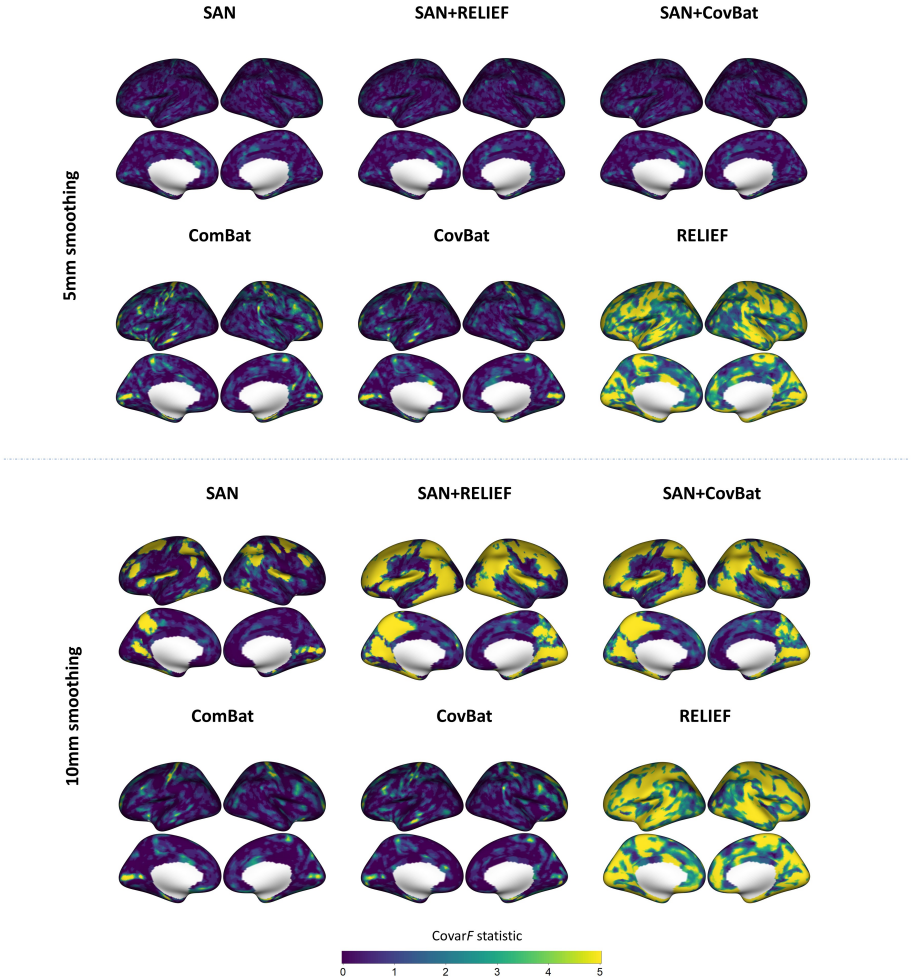

Figure 5: Covar$F$ statistic brain maps on smoothed data obtained from different harmonization methods.

15mm-20mm, 20mm-30mm and 30mm-40mm. With each of these ranges, we compute this measure using vertex pairs that fall within the specified range. For $\hat{\sigma}_{ij}^{(n)}(v, v^*)$ and $\hat{\sigma}_{i}^{(n)}(v, v^*)$, we apply harmonization methods to each bootstrap dataset and then calculate them as we do in section 3.3.2. For $\hat{\sigma}(v, v^*)$, we use the overall original dataset to estimate the underlying pooled covariances.

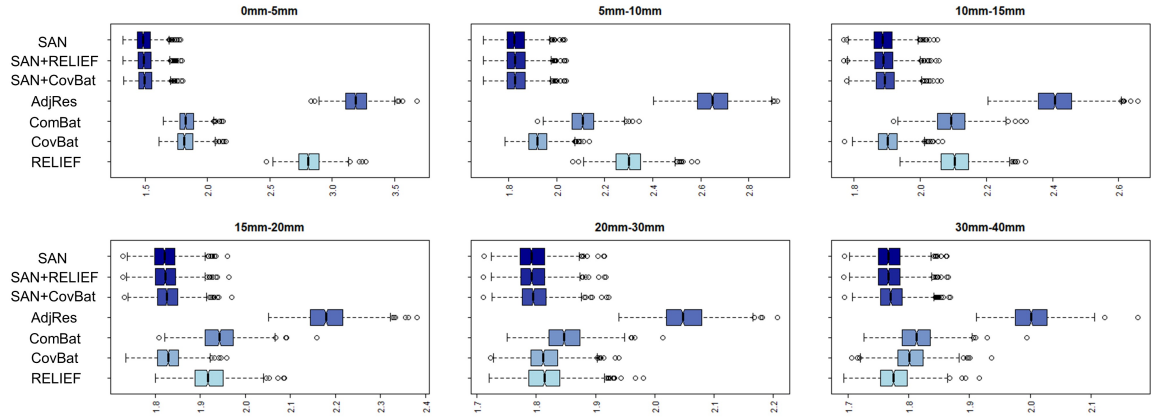

Figure 6: Boxplots illustrating Covar$F$ statistics for seven harmonization methods across different intervals in the Bootstrapping results.

Figure 6 shows boxplots of Covar$F$ statistics from bootstrapping, which align with the findings in Figure 2(b). Across the six ranges, SAN, SAN+RELIEF, and SAN+CovBat consistently exhibited smaller medians of Covar$F$ statistics compared to other methods, particularly for Covar$F$ statistics below 15mm. This can be attributed to their ability to capture and harmonize most of the scanner-specific local spatial dependencies in the data. Following them are CovBat and RELIEF. CovBat shows comparable performance to SAN in the 15mm-30mm range, while RELIEF shows similar Covar$F$ statistics to SAN in 30mm-40mm.

4. Discussion

In this paper, we propose a novel harmonization method, SAN, that identifies and parameterizes sources of heterogeneity in vertex-level cortical thickness data collected from different scanners or sites. We use Gaussian process to model and homogenize spatial covariances, and its probabilistic modeling ensures the smoothness of the harmonized cortical thickness data, which existing methods do not provide. SAN aligns with the growing need for providing high-quality data to downstream whole-brain analysis, especially those involving spatial covariance modelling. We use a two-stage approach to estimate scanner-specific parameters for the heterogeneous means and covariances, and use the method-of-moments estimators for scalability. SAN's flexible framework allows for integration with other covariance harmonization methods (e.g. CovBat, RELIEF) to further reduce potential latent scanner effects, although our data analysis suggests that spatial covariance explains most of the inter-scanner covariance heterogeneity of cortical thickness. Our analysis of the SPINS study suggests that there are specific anatomical regions that reveal a high degree of covariance heterogeneity across scanners, for which SAN's brain-level harmonization successfully homogenized. Although initially designed for vertex-wise data, SAN also proves advantageous for within-ROI and between-ROI analyses by ensuring scanner-specific spatial normalization at the foundational level, which is critical for constructing accurate ROI-level metrics.

In the analysis of vertex-level cortical thickness data from the SPINS study, we observe that a wide range of smoothing is discouraged before applying SAN (and other harmonization methods). In multi-site/scanner studies with inter-scanner biases in means and covariances, we provide empirical and theoretical evidence of higher mean and covariance heterogeneity induced by smoothing. Covariance harmonization methods (SAN, CovBat, RELIEF) appeared to be ineffective and performed worse than ComBat with highly smoothed data. Our sensitivity analysis of harmonization methods to varying levels of smoothing suggests that SAN is preferably applied to unsmoothed data or data with minimal smoothing levels (e.g., 5mm), but high smoothing levels (e.g., 10mm) negatively impact its performance. One possible explanation of this phenomenon is that spatial smoothing greatly increases the degree of spatial nonstationarity making it difficult for statistical harmonization methods to parametrize its sources appropriately. For these reasons, we recommend applying SAN to minimally smoothed (or unsmoothed) data, while more empirical and theoretical evidence is required to determine the optimal data processing pipeline.

SAN's harmonization aims to recover an optimal 'pooled' covariance from the parametrizations of the Gaussian process, which reduces the Covar$F$ statistic significantly than other harmonization methods we considered. However, a cautionary note is needed in choosing the harmonization method that meets users' research purposes. SAN is recommended when covariance modeling of the vertex-level cortical thickness data is critical in improving statistical inference. In contrast, RELIEF performs best in impeding the detection of scanners at the ROI level. However, RELIEF does not seem effective in reducing the Covar$F$ statistic as RELIEF primarily focuses on *removing* scanner-specific latent factors. These results suggest that, despite higher power shown in Zhang et al. [30] in massive univariate analysis, SAN would perform more promisingly in the spatial-extent inference that requires explicit spatial aucorrelation modeling.

SAN has room for improvement. First, we downsampled cortical thickness data to fsaverage5 space ($V \approx 10,000$) in our analysis, which seems to be sufficient in capturing dense spatial information without significant loss of information [42, 43]. However, SAN would be computationally limited when applied to higher resolutions (e.g., fsaverage6 or fsaverage7). Although implementing SAN is computationally feasible when $V \approx 10,000$, more research is needed to make it even more computationally efficient in higher dimensions. Second, longitudinal studies (e.g., Alzheimer's Disease Neuroimaging Initiatives) have identified inter-scanner biases in cortical thickness data [27]. Extending SAN to longitudinal neuroimaging studies would simultaneously account for both the within-subject variability and covariance heterogeneity across scanners. This enhancement could maintain within-subject dependencies in the harmonization process, thereby improving data quality for longitudinal designs not currently addressed by our existing SAN framework. Finally, there are also other structural imaging metrics based on the human cerebral cortex, such as surface area, and gyrification, which have shown discrepancies in their measurements across scanners [44, 45, 46]. Although this paper primarily focuses on harmonizing cortical thickness data, future research could

explore the validity of SAN in these imaging modalities.

5. Software

The R package for implementing SAN is publicly available at `https://github.com/junjypark/SAN`. Our harmonization took approximately 1 hour on a Macbook Air (M2,2022) with 16GB RAM to harmonize data with 9,354 imaging features from 357 subjects, which supports the computational feasibility of the proposed method. For server users, the harmonization was completed in approximately 36 minutes on a server cluster node (Lenovo SD350) equipped with 40 Intel "Skylake" cores (2.4GHz) and 202GB RAM. Parallel computing is supported by the package to mitigate computational costs working with a large number of scanners or sites.

Declaration of Competing Interests

None.

Acknowledgements

The SPINS study was supported by the National Institute of Mental Health (1/3R01MH102324-01, 2/3R01MH102313-01, 3/3R01MH102318-01). RZ was supported by the Doctoral Fellowship from the University of Toronto Data Science Institute. LDO was supported by the Brain & Behavior Research Foundation. ANV 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. JYP was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2022-04831), the University of Toronto's Data Science Institute (Catalyst Grant), McLaughlin Centre (Accelerator Grant), and the Connaught fund. The computing resources were enabled in part by support provided by University of Toronto and the Digital Research Alliance of Canada (`alliancecan.ca`).

Appendix A

For simplicity, suppose that a pair of imaging features is measured from two different scanners ($i = 1, 2,\ n_i = n$). We assume that $(y_{ij1}, y_{ij2})^\top$ follows spatial Gaussian process with mean zeros and scanner-specific variance-covariance structure,

$$\begin{pmatrix} y_{ij1} \\ y_{ij2} \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_i^2 + \sigma_i^2 & \sigma_i^2 \rho \\ \sigma_i^2 \rho & \tau_i^2 + \sigma_i^2 \end{bmatrix} \right),$$

where $0 < \rho < 1$ is analogous to $\exp(-\phi \cdot d)$ or $\exp(-\phi \cdot d^2)$ in SAN. If we apply smoothing, consider weights $w_1 > 0$ and $w_2 > 0$ that are $w_1 + w_2 = 1$. We have $y^s_{ij1} = w_1 y_{ij1} + w_2 y_{ij2}$ and $y^s_{ij2} = w_2 y_{ij1} + w_1 y_{ij2}$ where

$$\begin{pmatrix} y^s_{ij1} \\ y^s_{ij2} \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} (w_1^2 + w_2^2)(\tau_i^2 + \sigma_i^2) + 2w_1 w_2 \sigma_i^2 \rho & (w_1^2 + w_2^2)\sigma_i^2 \rho + 2w_1 w_2 (\tau_i^2 + \sigma_i^2) \\ (w_1^2 + w_2^2)\sigma_i^2 \rho + 2w_1 w_2 (\tau_i^2 + \sigma_i^2) & (w_1^2 + w_2^2)(\tau_i^2 + \sigma_i^2) + 2w_1 w_2 \sigma_i^2 \rho \end{bmatrix} \right).$$

To examine the covariance differences between scanners, we use the F statistic formula by plugging the true parameters:

$$\begin{aligned} F &= \frac{[E(y_{1j1}y_{1j2}) - E(y_{2j1}y_{2j2})]^2}{[Var(y_{1j1}y_{1j2}) + Var(y_{2j1}y_{2j2})]/n} \\ &= \frac{(\sigma_1^2\rho - \sigma_2^2\rho)^2}{[(\tau_1^2 + \sigma_1^2)^2 + \sigma_1^4\rho^2 + (\tau_2^2 + \sigma_2^2)^2 + \sigma_2^4\rho^2]/n}, \end{aligned}$$

which follows from $E(y_{ij1}y_{ij2}) = \sigma_i^2\rho$ and $Var(y_{ij2}y_{ij2}) = (\tau_i^2 + \sigma_i^2)^2 + \sigma_i^4\rho^2$. Similarly, the F statistic for the smoothed data, denoted by F^s, is

$$\begin{aligned} F^s &= \frac{[E(y^s_{1j1}y^s_{1j2}) - E(y^s_{2j1}y^s_{2j2})]^2}{(Var(y^s_{1j1}y^s_{1j2}) + Var(y^s_{2j1}y^s_{2j2}))/n} \\ &= \frac{[\sigma_1^2\rho + w(\tau_1^2 + \sigma_1^2) - (\sigma_2^2\rho + w(\tau_2^2 + \sigma_2^2))]^2}{[((\tau_1^2 + \sigma_1^2) + w\sigma_1^2\rho)^2 + (\sigma_1^2\rho + w(\tau_1^2 + \sigma_1^2))^2 + ((\tau_2^2 + \sigma_2^2) + w\sigma_2^2\rho)^2 + (\sigma_2^2\rho + w(\tau_2^2 + \sigma_2^2))^2]/n}, \end{aligned}$$

where $w = \frac{2w_1w_2}{w_1^2+w_2^2}$. If smoothed features show a larger covariance heterogeneity between scanners than unsmoothed features, the statistic F^s of smoothed features will be larger than the unsmoothed F. When τ_i and σ_i^2 are monotonic (e.g., $\tau_1^2 > \tau_2^2$, then $\sigma_1^2 > \sigma_2^2$) we have $F^s > F$ because $F^s/F = A \times B$ where

$$\begin{aligned} A &= \left(w_1^2 + w_2^2 + \frac{2w_1w_2}{\rho}\left(1 + \frac{\tau_1^2 - \tau_2^2}{\sigma_1^2 - \sigma_2^2}\right)\right)^2 \\ B &= \frac{(\tau_1^2 + \sigma_1^2)^2 + \sigma_1^4\rho^2 + (\tau_2^2 + \sigma_2^2)^2 + \sigma_2^4\rho^2}{(w_1^2 + w_2^2)[((\tau_1^2 + \sigma_1^2) + w\sigma_1^2\rho)^2 + (\sigma_1^2\rho + w(\tau_1^2 + \sigma_1^2))^2 + ((\tau_2^2 + \sigma_2^2) + w\sigma_2^2\rho)^2 + (\sigma_2^2\rho + w(\tau_2^2 + \sigma_2^2))^2]}. \end{aligned}$$

Here, $A > 1$ follows from $2w_1w_2 < \frac{2w_1w_2}{\rho}\left(1 + \frac{\tau_1^2-\tau_2^2}{\sigma_1^2-\sigma_2^2}\right)$, and $B > 1$ is derived by using

$$((w_1^2 + w_2^2)(\tau_i^2 + \sigma_i^2) + 2w_1w_2\sigma_i^2\rho)^2 + ((w_1^2 + w_2^2)\sigma_i^2\rho + 2w_1w_2(\tau_i^2 + \sigma_i^2))^2 < (\tau_i^2 + \sigma_i^2)^2 + \sigma_i^4\rho^2.$$

References

[1] M. Goto, O. Abe, A. Hagiwara, S. Fujita, K. Kamagata, M. Hori, S. Aoki, T. Osada, S. Konishi, Y. Masutani, et al., Advantages of using both voxel-and surface-based morphometry in cortical

morphology analysis: a review of various applications, Magnetic Resonance in Medical Sciences 21 (2022) 41–57.

[2] M. Thambisetty, J. Wan, A. Carass, Y. An, J. L. Prince, S. M. Resnick, Longitudinal changes in cortical thickness associated with normal aging, Neuroimage 52 (2010) 1215–1223.

[3] E. R. Sowell, B. S. Peterson, P. M. Thompson, S. E. Welcome, A. L. Henkenius, A. W. Toga, Mapping cortical change across the human life span, Nature neuroscience 6 (2003) 309–315.

[4] H. Cho, S. Jeon, S. J. Kang, J.-M. Lee, J.-H. Lee, G. H. Kim, J. S. Shin, C. H. Kim, Y. Noh, K. Im, et al., Longitudinal changes of cortical thickness in early-versus late-onset alzheimer's disease, Neurobiology of aging 34 (2013) 1921–e9.

[5] J. P. Lerch, J. C. Pruessner, A. Zijdenbos, H. Hampel, S. J. Teipel, A. C. Evans, Focal decline of cortical thickness in alzheimer's disease identified by computational neuroanatomy, Cerebral cortex 15 (2005) 995–1001.

[6] N. E. Van Haren, H. G. Schnack, W. Cahn, M. P. Van Den Heuvel, C. Lepage, L. Collins, A. C. Evans, H. E. H. Pol, R. S. Kahn, Changes in cortical thickness during the course of illness in schizophrenia, Archives of general psychiatry 68 (2011) 871–880.

[7] K. L. Narr, R. M. Bilder, A. W. Toga, R. P. Woods, D. E. Rex, P. R. Szeszko, D. Robinson, S. Sevy, H. Gunduz-Bruce, Y.-P. Wang, et al., Mapping cortical thickness and gray matter concentration in first episode schizophrenia, Cerebral cortex 15 (2005) 708–719.

[8] M. Sailer, B. Fischl, D. Salat, C. Tempelmann, M. A. SchoÈnfeld, E. Busa, N. Bodammer, H.-J. Heinze, A. Dale, Focal thinning of the cerebral cortex in multiple sclerosis, Brain 126 (2003) 1734–1744.

[9] C. Tsagkas, M. M. Chakravarty, L. Gaetano, Y. Naegelin, M. Amann, K. Parmar, A. Papadopoulou, J. Wuerfel, L. Kappos, T. Sprenger, et al., Longitudinal patterns of cortical thinning in multiple sclerosis, Human brain mapping 41 (2020) 2198–2215.

[10] L. C. Hanford, A. Nazarov, G. B. Hall, R. B. Sassi, Cortical thickness in bipolar disorder: a systematic review, Bipolar disorders 18 (2016) 4–18.

[11] S. Li, R. Bai, Y. Yang, R. Zhao, B. Upreti, X. Wang, S. Liu, Y. Cheng, J. Xu, Abnormal cortical thickness and structural covariance networks in systemic lupus erythematosus patients without major neuropsychiatric manifestations, Arthritis Research & Therapy 24 (2022) 1–19.

[12] J. L. Bernal-Rusiel, D. N. Greve, M. Reuter, B. Fischl, M. R. Sabuncu, A. D. N. Initiative, et al., Statistical analysis of longitudinal neuroimage data with linear mixed effects models, Neuroimage 66 (2013) 249–260.

[13] T. Ge, T. E. Nichols, P. H. Lee, A. J. Holmes, J. L. Roffman, R. L. Buckner, M. R. Sabuncu, J. W. Smoller, Massively expedited genome-wide heritability analysis (megha), Proceedings of the National Academy of Sciences 112 (2015) 2479–2484.

[14] M. A. Lindquist, A. Mejia, Zen and the art of multiple comparisons, Psychosomatic medicine 77 (2015) 114.

[15] J. L. Bernal-Rusiel, M. Reuter, D. N. Greve, B. Fischl, M. R. Sabuncu, A. D. N. Initiative, et al., Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data, Neuroimage 81 (2013) 358–370.

[16] J. Y. Park, M. Fiecas, A. D. N. Initiative, et al., Permutation-based inference for spatially localized signals in longitudinal mri data, NeuroImage 239 (2021) 118312.

[17] S. M. Weinstein, S. N. Vandekar, E. B. Baller, D. Tu, A. Adebimpe, T. M. Tapera, R. C. Gur, R. E. Gur, J. A. Detre, A. Raznahan, et al., Spatially-enhanced clusterwise inference for testing and localizing intermodal correspondence, NeuroImage 264 (2022) 119712.

[18] J. Y. Park, M. Fiecas, Clean: Leveraging spatial autocorrelation in neuroimaging data in clusterwise inference, Neuroimage 255 (2022) 119192.

[19] R. Pan, E. W. Dickie, C. Hawco, N. Reid, A. N. Voineskos, J. Y. Park, Spatial-extent inference for testing variance components in reliability and heritability studies, bioRxiv (2023) 2023–04.

[20] X. Han, J. Jovicich, D. Salat, A. van der Kouwe, B. Quinn, S. Czanner, E. Busa, J. Pacheco, M. Albert, R. Killiany, et al., Reliability of mri-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer, Neuroimage 32 (2006) 180–194.

[21] H. G. Schnack, N. E. van Haren, R. M. Brouwer, G. C. M. van Baal, M. Picchioni, M. Weisbrod, H. Sauer, T. D. Cannon, M. Huttunen, C. Lepage, et al., Mapping reliability in multicenter mri: Voxel-based morphometry and cortical thickness, Human brain mapping 31 (2010) 1967–1982.

[22] J. Jovicich, S. Czanner, D. Greve, E. Haley, A. van Der Kouwe, R. Gollub, D. Kennedy, F. Schmitt, G. Brown, J. MacFall, et al., Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data, Neuroimage 30 (2006) 436–443.

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

[24] J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur, et al., Harmonization of multi-site diffusion tensor imaging data, NeuroImage 161 (2017) 149–170. doi:https://doi.org/10.1016/j.neuroimage.2017.08.047.

[25] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, et al., Harmonization of cortical thickness measurements across scanners and sites, NeuroImage 167 (2018) 104–120.

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

[27] J. C. Beer, N. J. Tustison, P. A. Cook, C. Davatzikos, Y. I. Sheline, R. T. Shinohara, K. A. Linn, A. D. N. Initiative, et al., Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data, Neuroimage 220 (2020) 117129.

[28] C.-L. Chen, M. E. Torbati, J. D. Wilson, D. S. Minhas, C. M. Laymon, S. J. Hwang, P. Maillard, E. Fletcher, C. DeCarli, D. Tudorascu, Reducing mri inter-scanner variability using 3d superpixel combat, in: Alzheimer's Association International Conference, ALZ, 2022.

[29] A. A. Chen, J. C. Beer, N. J. Tustison, P. A. Cook, R. T. Shinohara, H. Shou, A. D. N. Initiative, Mitigating site effects in covariance for machine learning in neuroimaging data, Human Brain Mapping 43 (2022) 1179–1195.

[30] R. Zhang, L. D. Oliver, A. N. Voineskos, J. Y. Park, RELIEF: A structured multivariate approach for removal of latent inter-scanner effects, Imaging Neuroscience 1 (2023) 1–16. URL: https://doi.org/10.1162/imag_a_00011. doi:10.1162/imag_a_00011. arXiv:https://direct.mit.edu/imag/article-pdf/doi/10.1162/imag_a_00011/2156061/imag_a_00011.

[31] R. W. Cox, G. Chen, D. R. Glen, R. C. Reynolds, P. A. Taylor, Fmri clustering in afni: false-positive rates redux, Brain connectivity 7 (2017) 152–171.

[32] G. Wang, J. Muschelli, M. A. Lindquist, Moderated t-tests for group-level fmri analysis, NeuroImage 237 (2021) 118141.

[33] T. Zou, W. Lan, H. Wang, C.-L. Tsai, Covariance regression analysis, Journal of the American Statistical Association 112 (2017) 266–281.

[34] J. A. Nelder, R. Mead, A Simplex Method for Function Minimization, The Computer Journal 7 (1965) 308–313. URL: https:

//doi.org/10.1093/comjnl/7.4.308. doi:10.1093/comjnl/7.4.308. arXiv:https://academic.oup.com/comjnl/article-pdf/7/4/308/1013182/7-4-308.pdf.

[35] J. D. Viviano, R. W. Buchanan, N. Calarco, J. M. Gold, G. Foussias, N. Bhagwat, L. Stefanik, C. Hawco, P. DeRosse, M. Argyelan, et al., Resting-state connectivity biomarkers of cognitive performance and social function in individuals with schizophrenia spectrum disorder and healthy control subjects, Biological Psychiatry 84 (2018) 665–674.

[36] L. D. Oliver, C. Hawco, P. Homan, J. Lee, M. F. Green, J. M. Gold, P. DeRosse, M. Argyelan, A. K. Malhotra, R. W. Buchanan, et al., Social cognitive networks and social cognitive performance across individuals with schizophrenia spectrum disorders and healthy control participants, Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 6 (2021) 1202–1214.

[37] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, et al., fmriprep: a robust preprocessing pipeline for functional mri, Nature methods 16 (2019) 111–116.

[38] K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, S. S. Ghosh, Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python, Frontiers in neuroinformatics 5 (2011) 13.

[39] B. B. Avants, C. L. Epstein, M. Grossman, J. C. Gee, Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain, Medical image analysis 12 (2008) 26–41.

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

[41] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al., An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest, Neuroimage 31 (2006) 968–980.

[42] A. F. Mejia, Y. R. Yue, D. Bolin, F. Lindgren, M. A. Lindquist, A bayesian general linear modeling approach to cortical surface fmri data analysis, Journal of the American Statistical Association (2019).

[43] T. Xu, A. Opitz, R. C. Craddock, M. J. Wright, X.-N. Zuo, M. P. Milham, Assessing variations in areal organization for the intrinsic brain: from fingerprints to reliability, Cerebral Cortex 26 (2016) 4192–4211.

[44] E. Iannopollo, K. Garcia, A. N. Initiative, Enhanced detection of cortical atrophy in alzheimer's disease using structural mri with anatomically constrained longitudinal registration, Human Brain Mapping 42 (2021) 3576–3592.

[45] F. H. de Moraes, V. B. Mello, F. Tovar-Moll, B. Mota, Establishing a baseline for human cortical folding morphological variables: a multisite study, Frontiers in Neuroscience 16 (2022) 897226.

[46] J. Radua, E. Vieta, R. Shinohara, P. Kochunov, Y. Quidé, M. J. Green, C. S. Weickert, T. Weickert, J. Bruggemann, T. Kircher, et al., Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma, Neuroimage 218 (2020) 116956.