A Method for Harmonizing Longitudinal Multi-scanner Imaging Data

Joanne C. Beer[a,*], Nicholas J. Tustison[b], Philip A. Cook[c,d], Yvette I. Sheline[c,d,e], Russell T. Shinohara[a,d], Kristin A. Linn[a,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] *Penn Statistics in Imaging and Visualization Endeavor (PennSIVE) Center, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, United States*
[b] *Department of Radiology and Medical Imaging, University of Virginia, United States*
[c] *Center for Neuromodulation in Depression and Stress, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, United States*
[d] *Department of Radiology, Perelman School of Medicine, University of Pennsylvania, United States*
[e] *Department of Neurology, Perelman School of Medicine, University of Pennsylvania, United States*

Abstract

While aggregation of neuroimaging datasets from multiple sites and scanners can yield increased statistical power, it also presents challenges due to systematic scanner effects. This unwanted technical variability can introduce noise and bias into estimation of biological variability of interest. We propose a method for harmonizing longitudinal multi-scanner imaging data based on ComBat, a method originally developed for genomics and later adapted to cross-sectional neuroimaging data. Using longitudinal cortical thickness measurements from 663 participants in the Alzheimer's Disease Neuroimaging

*Corresponding author

Email addresses: joanne.beer@pennmedicine.upenn.edu (Joanne C. Beer), klinn@pennmedicine.upenn.edu (Kristin A. Linn)

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Initiative (ADNI) study, we demonstrate the presence of additive and multiplicative scanner effects in various brain regions. We compare estimates of the association between diagnosis group and change in cortical thickness over time using three versions of the ADNI data: unharmonized data, data harmonized using cross-sectional ComBat, and data harmonized using longitudinal ComBat. In simulation studies, we show that longitudinal ComBat is more powerful for detecting longitudinal change than cross-sectional ComBat, and controls the type I error rate better than unharmonized data with scanner included as a covariate.

Keywords: ADNI, Alzheimer's, ComBat, cortical thickness, harmonization, MRI

---

1. Introduction

Aggregation of neuroimaging data across sites and scanners can potentially increase statistical power to detect biological variability of interest. However, the use of different scanner hardware, software, and acquisition protocols can introduce unwanted technical variability (Han et al., 2006; Jovicich et al., 2006; Takao et al., 2011). ComBat (named for "combining batches") is an empirical Bayesian method for data harmonization that was originally designed for genomics (Johnson et al., 2007). It has recently been adapted to neuroimaging studies and applied to diverse data types including diffusion tensor imaging (DTI) (Fortin et al., 2017), cortical thicknesses (Fortin et al., 2018), functional connectivity measurements (Yu et al., 2018), and radiomic features derived from positron emission tomography (PET) imaging (Orlhac et al., 2018). In general, ComBat is applicable to situations where multiple features of the same type are measured for each participant, where features might be expression levels for different genes, or imaging-derived metrics from different voxels or anatomic regions. In this paper, we extend the ComBat methodology from a cross-sectional to a longitudinal setting, where participants are imaged repeatedly over the course of the study.

In contrast to a general linear model approach that includes site or scanner as a fixed effect covariate, there are several benefits to the empirical Bayes estimation method used in ComBat. Notably, ComBat is more robust to outliers in the case of small within-scanner sample sizes (Johnson et al., 2007). ComBat assumes that for a given scanner, the scanner effects across features derive from a common distribution, and thus borrows information

across features to shrink estimates towards a common mean. Furthermore, in addition to removing additive scanner effects, ComBat also corrects multiplicative scanner effects by removing heteroscedasticity of model errors across scanners. Prior studies have shown that the location (mean) and scale (variance) adjustment implemented in ComBat outperforms methods that merely include scanner as a covariate (Fortin et al., 2018).

While longitudinal studies are important for measuring within-subject change, there has been little work on longitudinal data harmonization. Müller et al. (2016) examined a variety of batch correction methods in longitudinal gene expression data and found that a combination of quantile normalization and ComBat performed best. However, their batch effect estimation method relied on biological replicates processed at both baseline and follow-up. This does not translate well to longitudinal neuroimaging study designs, as there is no way to obtain analogous biological replicates. Venkatraman et al. (2015) estimated scanner fixed effects in cross-sectional and longitudinal DTI data using linear mixed effects models. They then used these estimates to apply a linear correction to new data. The authors found that accounting for within-subject variability led to better scanner effect estimates in longitudinal as compared with cross-sectional data. However, their method does not enjoy the benefits of empirical Bayes discussed above, nor does it adjust for multiplicative scanner effects.

In this work, we aim to estimate and correct for additive and multiplicative scanner effects while explicitly accounting for the within-subject correlation inherent to longitudinal studies, such that the harmonization method may be flexibly applied to existing and future longitudinal multi-scanner neuroimaging datasets. We illustrate the longitudinal ComBat method using cortical thickness data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Weiner et al., 2015). Alzheimer's disease (AD) is a neurodegenerative disease characterized by aggregation of amyloid β plaques and accumulation of neurofibrillary tangles. Brain atrophy is one of the earliest biomarkers of AD that is visible on structural magnetic resonance imaging (MRI), particularly in certain regions such as the hippocampus and entorhinal cortex (Dickerson et al., 2008; Bakkour et al., 2009). ADNI is a multi-site longitudinal study including cognitively normal, mild cognitive impairment (MCI), which is a prodromal stage of AD, and AD participants.

In Section 2 we describe the ADNI data and assess the presence of scanner effects; in Section 3 we outline the proposed longitudinal ComBat harmonization method; in Section 4 we use the ADNI dataset to compare model

estimates and inference for longitudinal ComBat, cross-sectional ComBat, and unharmonized data; and in Section 5 we present a simulation study. We provide discussion and conclusions in Section 6. Code for implementing longitudinal ComBat is available at `https://github.com/jcbeer/longCombat`.

2. Quantifying site and scanner effects in ADNI data

2.1. Methods

We examined longitudinal cortical thickness data from participants enrolled in the first phase of the ADNI study. Data were obtained from the ADNI database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see `www.adni-info.org`.

We included 663 ADNI-1 participants from 58 study sites. (See Figure 1A for distributions of participant age, sex, and diagnoses.) Structural MRI brain scans were done at 6 or 12 month intervals for up to 3 years from baseline. Many sites used multiple MRI scanners over the course of the study, and a given participant may have been scanned on different scanners across visits. The data was acquired on 142 total scanners, of which 41 were 3.0 T and the remainder were 1.5 T. Participants were diagnosed at baseline as cognitively normal (CN, $n = 197$), late mild cognitive impairment (LMCI, $n = 324$), or Alzheimer's disease (AD, $n = 142$). They were reassessed at each study visit, but no participants changed diagnostic category during the study.

Cortical thicknesses for 62 brain regions defined using the Desikan-Killiany-Tourville atlas (Klein and Tourville, 2012) were obtained using the Advanced Normalization Tools (ANTs) longitudinal cortical thickness pipeline (Tustison et al., 2019). Specifically, we used data processed with the ANTs Longitudinal-SST pipeline, which involves first rigidly transforming each subject to a single subject template (SST) and then estimating cortical thickness in the SST space. In comparison to the well-known FreeSurfer longitudinal processing pipeline, ANTs Longitudinal-SST results in superior statistical power for differentiation of diagnostic groups in this dataset, with greater between-subject to residual variance ratios and tighter confidence and prediction intervals (Tustison et al., 2019). For further details on this dataset,

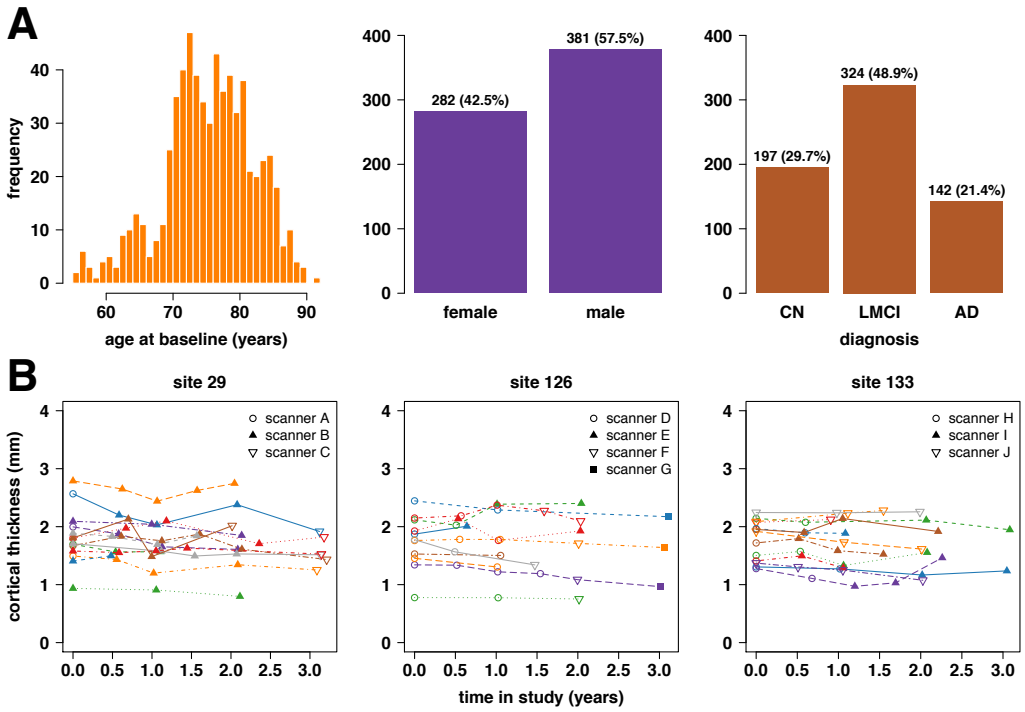

Figure 1: (A) Characteristics of $n = 663$ ADNI-1 participants. (B) Example trajectories for left superior frontal cortical thickness at 3 ADNI sites.

please refer to Tustison et al. (2019) and references therein. Sample trajectories of the unharmonized cortical thickness data are depicted in Figure 1B.

We first assessed via statistical testing whether site or scanner additive (i.e., shift in mean) and multiplicative effects (i.e., heteroscedasticity) were present in the data, while also controlling for known differences in biological variability (i.e., age, sex, diagnosis) across site or scanner. We considered two sources of potential technical variability (encoded as the 'scanner' variable in the model below): (1) site effect only ($m = 58$), (2) scanner effect, where scanners were identified as unique combinations of site, scanner vendor, model, head coil, and field strength variables ($m = 126$, 16 images omitted because $n_i = 1$, where n_i denotes total number of images obtained from scanner i). For both of these scenarios and for each of the $V = 62$

cortical regions, we fit the linear mixed effects model

$$
\begin{aligned}
y_{ij\nu}(t_k) = \alpha_\nu &+ \beta_{\nu 1}(\text{baseline age}) + \beta_{\nu 2} I(\text{sex} = \text{male}) \\
&+ \beta_{\nu 3} I(\text{diagnosis} = \text{LMCI}) + \beta_{\nu 4} I(\text{diagnosis} = \text{AD}) \\
&+ \beta_{\nu 5} \cdot t_k + \beta_{\nu 6} I(\text{diagnosis} = \text{LMCI}) \cdot t_k + \beta_{\nu 7} I(\text{diagnosis} = \text{AD}) \cdot t_k \\
&+ \beta_{\nu 8} I(\text{scanner}_k = 2) + \cdots + \beta_{\nu(m+7)} I(\text{scanner}_k = m) \\
&+ \eta_{j\nu} + \epsilon_{j\nu}(t_k),
\end{aligned} \tag{1}
$$

where $i \in \{1, \ldots, m\}$ is the site or scanner index, $j \in \{1, \ldots, N\}$ is the participant index, $\nu \in \{1, \ldots V\}$ is the feature index (corresponding to the 62 regional cortical thickness measurements, in this case), $k \in \{0, \ldots, K_j\}$ is the visit index and K_j is total number of visits for participant j, $t_k \in \mathbb{R}_{\geq 0}$ is years from baseline visit for visit k, α_ν is an intercept term, $I(\cdot)$ is an indicator function equal to one if the argument condition is true and zero otherwise. Reference levels for factor variables are female sex, cognitively normal diagnosis, and $\text{scanner}_k = 1$. All parameters represent fixed effects except for the subject-specific random intercept, $\eta_{j\nu}$, for which we assume the distribution $N(0, \rho_\nu)$, and the error term, $\epsilon_{j\nu}(t_k)$, for which we assume the distribution $N(0, \sigma_\nu)$.

Models were fit using the R package `lme4` (Bates et al., 2015) and R version 3.5.3 (R Core Team, 2019). To test for additive site or scanner effects, we also fit models omitting the site or scanner fixed effects and used the package `pbkrtest` (Halekoh and Højsgaard, 2014) to carry out tests of their joint significance using the Kenward-Roger (KR) approach (Kenward and Roger, 1997). We also tested for a differential scaling effect by site or scanner. We fit the model represented in Equation (1) above, including the site or scanner fixed effects. Due to small within-site or within-scanner sample sizes in some cases, we used the non-parametric Fligner-Killeen (FK) test (Conover et al., 1981) to assess heteroscedasticity of the residuals ($\widehat{\epsilon}_{j\nu}(t)$) across site or scanner. Additionally, we tested whether incorporating specific scanner information rather than site alone significantly improved the model. Since the two scenarios correspond to nested models, we used the KR test. Finally, we did exploratory visualizations to assess whether additive and multiplicative scanner effects were associated with scanner field strength, vendor, number of subjects scanned, total number of scans, percentage of scans with AD diagnosis, or percentage of scans with CN diagnosis.

All brain figures in this manuscript were made using `freesurfer_statsurf_display` (Murdoch Childrens Research Institute Developmental Imaging Group, 2017)

and MATLAB R2018a (MATLAB, 2018).

2.2. Results

The KR test for additive site effects was significant ($p < 0.05$) for all but 8 of the 62 regional cortical thickness measurements (i.e., "features"; Supplementary Figure S1 and Table S1). The KR test for additive scanner effects was significant for all features (Supplementary Figure S2 and Table S2). Figure 2A illustrates the additive scanner effects for the feature with the largest KR test F-statistic and shows $-\log_{10} p$-values across brain regions. Additive scanner effects were particularly large in medial occipital and medial temporal regions. Visualizations showed that 3.0 T scanners tended to result in larger cortical thickness measurements (Supplementary Figures S5-S10).

Across both site and scanner, all features had significantly different residual variances (Supplementary Figures S3 and S4, Supplementary Tables S3 and S4). Figure 2B illustrates the multiplicative scanner effects for the feature with the largest FK test χ^2-statistic and shows $-\log_{10} p$-values across brain regions. Multiplicative scanner effects were particularly prominent in superior frontal and superior parietal regions. Visualizations indicated that vendor 1 scanners generally tended to have larger, while vendor 3 scanners had smaller, residual variability, with vendor 2 scanners falling in between (Supplementary Figures S11-S16).

Scanner significantly improved the model for all 62 features (Supplementary Table S5). Hence, we use scanner instead of site in all subsequent analyses.

3. Longitudinal ComBat

3.1. Longitudinal ComBat model

For a longitudinal version of the ComBat harmonization method, we propose the model

$$y_{ij\nu}(t) = \alpha_\nu + \gamma_{i\nu} + \mathbf{X}_j^T(t)\boldsymbol{\beta}_\nu + \eta_{j\nu} + \delta_{i\nu}\epsilon_{ij\nu}(t),$$

where $i \in \{1, \ldots, m\}$ is the scanner index, $j \in \{1, \ldots, N\}$ is the participant index, $\nu \in \{1, \ldots V\}$ is the feature index, t is time (continuous or categorical), $y_{ij\nu}(t)$ is the observed data for feature ν, participant j, scanner i, and time t, α_ν is overall mean for feature ν at baseline, $\gamma_{i\nu}$ is the additive scanner i parameter for feature ν, $\mathbf{X}_j(t)$ is a $p \times 1$ vector of potentially time-varying

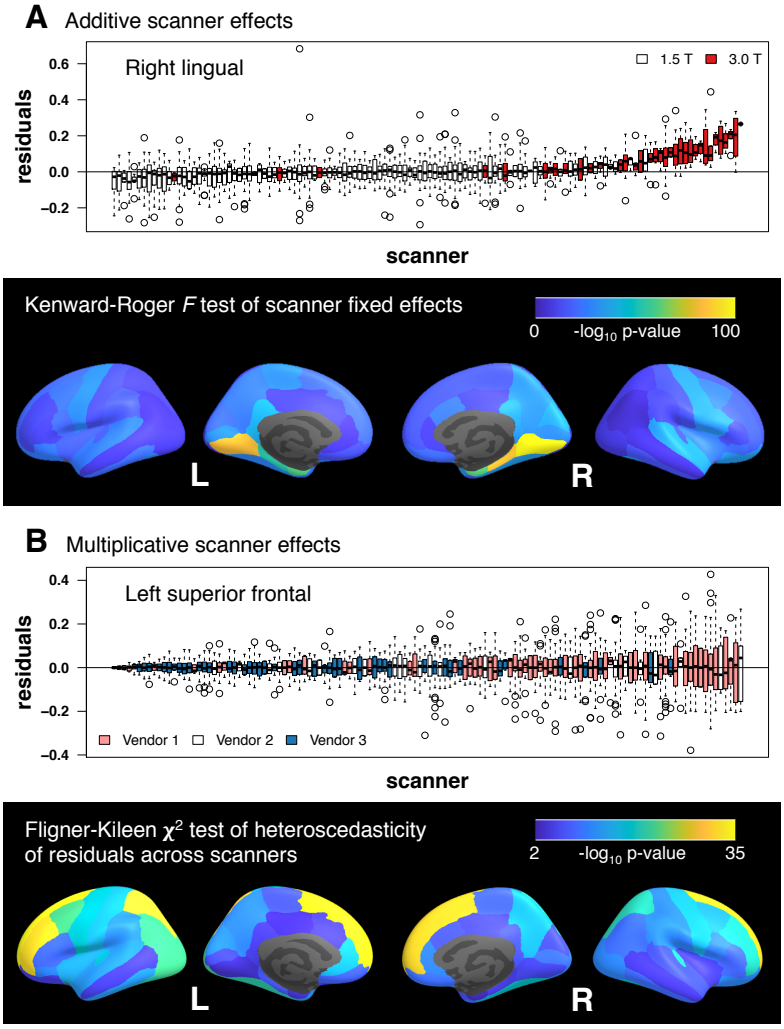

Figure 2: (A) Additive scanner effects. Boxplots show distributions of residuals across scanners after fitting a model with baseline age, sex, diagnosis, time, and diagnosis $\times$ time fixed effects and a subject-specific random intercept. Right lingual cortex was the region with the largest additive scanner effects according to the Kenward-Roger F-test; parahippocampal and entorhinal cortical regions also showed large effects. 3.0 T scanners tended to produce larger estimates of cortical thickness than 1.5 T scanners. (B) Multiplicative scanner effects. Boxplots show distributions of residuals across scanners after fitting a model with baseline age, sex, diagnosis, time, scanner, and diagnosis $\times$ time fixed effects and a subject-specific random intercept. Left superior frontal cortex was the region with the largest multiplicative scanner effects according to the Fligner-Kileen χ^2-test. Vendor 1 scanners tended to have larger, while vendor 3 scanners had smaller, residual variability.

covariates for participant j at time t (e.g., age, sex, the outcome that we ultimately intend to assess in association with the harmonized data such as diagnosis group or cognitive test score, and time), $\boldsymbol{\beta}_\nu$ is a $p \times 1$ vector of coefficients for feature ν, $\eta_{j\nu}$ is a subject-specific random intercept for participant j and feature ν, $\delta_{i\nu}$ is the scanner i scaling factor for feature ν, and $\epsilon_{j\nu}(t)$ is the error term. We assume $\eta_{ij\nu} \sim N(0, \rho_\nu^2)$ and $\epsilon_{j\nu}(t) \sim N(0, \sigma_\nu^2)$.

The ComBat-harmonized data is

$$y_{ij\nu}^{\text{ComBat}}(t) = \frac{y_{ij\nu}(t) - \widehat{\alpha}_\nu - \widehat{\gamma}_{i\nu} - \mathbf{X}_j^T(t)\widehat{\boldsymbol{\beta}}_\nu - \widehat{\eta}_{j\nu}}{\widehat{\delta}_{i\nu}} + \widehat{\alpha}_\nu + \mathbf{X}_j^T(t)\widehat{\boldsymbol{\beta}}_\nu + \widehat{\eta}_{j\nu},$$

where $\widehat{\alpha}_\nu, \widehat{\gamma}_{i\nu}, \widehat{\boldsymbol{\beta}}_\nu, \widehat{\eta}_{j\nu}$, and $\widehat{\delta}_{i\nu}$ are parameter estimates.

3.2. Parameter estimation

3.2.1. Standardization step

The empirical Bayes estimation for ComBat parameters assumes that for a given scanner, the additive scanner parameters across features ν all derive from a common distribution, $\gamma_{i\nu} \sim N(\gamma_i, \tau_i^2)$, and similarly for scanner scaling factors, $\delta_{i\nu}^2 \sim$ Inverse Gamma(λ_i, θ_i). To obtain unbiased empirical Bayes prior distribution estimates of scanner effects, we first standardize features so they have similar overall mean and variance. For this step, Johnson et al. (2007) used a feature-wise ordinary least squares approach to obtain the estimates $\widehat{\alpha}_\nu, \widehat{\boldsymbol{\beta}}_\nu, \widehat{\gamma}_{i\nu}$. To properly account for the dependence of repeated within-subject observations, we propose using a feature-wise linear mixed effects model with a random subject-specific intercept, $\eta_{j\nu} \sim N(0, \rho_\nu^2)$. We estimate the fixed effect parameters $\alpha_\nu, \boldsymbol{\beta}_\nu, \gamma_{i\nu}$ using the best linear unbiased estimator (BLUE), the subject effect variance ρ_ν^2 and error variance σ_ν^2 with the restricted maximum likelihood (REML) estimator, and subject-specific intercepts $\eta_{j\nu}$ using the best linear unbiased predictor (BLUP). For parameter identifiability, we constrain $\sum_i n_i \widehat{\gamma}_{i\nu}^{BLUE} = 0$, where n_i the is total number of images from scanner i.

Standardized data are calculated as

$$z_{ij\nu}(t) = \frac{y_{ij\nu}(t) - \widehat{\alpha}_\nu^{BLUE} - \mathbf{X}_j^T(t)\widehat{\boldsymbol{\beta}}_\nu^{BLUE} - \widehat{\eta}_{j\nu}^{BLUP}}{\widehat{\sigma}_\nu^{REML}}.$$

We assume that the standardized data $z_{ij\nu}(t)$ are from the distribution $N(\gamma_{i\nu}, \delta_{i\nu}^2)$. Prior distributions on the scanner effect parameters are assumed to be $\gamma_{i\nu} \sim N(\gamma_i, \tau_i^2)$, and $\delta_{i\nu}^2 \sim$ Inverse Gamma(λ_i, θ_i).

3.2.2. Empirical Bayes estimation of scanner effects

After standardization, parameter estimation for longitudinal ComBat is similar to that for standard ComBat. Hyperparameters $\gamma_i, \tau_i^2, \lambda_i, \theta_i$ are estimated from standardized data using the method of moments, and empirical Bayes estimates for scanner effect parameters $\gamma_{i\nu}$ and $\delta_{i\nu}^2$ are given by conditional posterior means. Please refer to Appendix A for derivations of these estimators.

3.3. Longitudinal ComBat-harmonized data

Finally, we use the empirical Bayes estimates $\widehat{\gamma}_{i\nu}^*$ and $\widehat{\delta}_{i\nu}^{2*}$ and the linear mixed effects model estimates to adjust the data:

$$y_{ij\nu}^{\text{ComBat}}(t) = \frac{\widehat{\sigma}_\nu^{REML}}{\widehat{\delta}_{i\nu}^*}\left(z_{ij\nu}(t) - \widehat{\gamma}_{i\nu}^*\right) + \widehat{\alpha}_\nu^{BLUE} + \mathbf{X}_j^T(t)\widehat{\boldsymbol{\beta}}_\nu^{BLUE} + \widehat{\eta}_{j\nu}^{BLUP}.$$

The first term performs the location and scale adjustment, thereby removing additive and multiplicative scanner effects, and remultiplies by the residual variance estimate to put features back on their original scale; the remaining terms add back estimates of biological effects of interest.

4. Comparison of data harmonization approaches in ADNI data

4.1. Methods

We used the ADNI cortical thickness data to compare three data harmonization approaches: (1) data harmonized across scanners using the proposed longitudinal ComBat method, (2) data harmonized across scanners using a cross-sectional version of ComBat, which does not account for the within-subject repeated measures (i.e., we omit the subject-specific random intercept from the ComBat steps, but include it in the final model), and (3) unharmonized cortical thickness data. Since our method is specifically designed for longitudinal data, we focused on the diagnosis by time interaction coefficients. Parameters of interest are $\beta_{\nu 6}$ and $\beta_{\nu 7}$ in Equation (1); these quantify differential rates of cortical thickness loss over time for LMCI and AD groups, respectively, relative to CN. For each of the three harmonization approaches, we fit the model given in Equation (1) for each feature, using the corresponding harmonized or unharmonized outcomes, and either included or omitted the scanner fixed effects.

We evaluated the results using the following criteria. As in Section 2 for the unharmonized data, we first tested whether any residual additive or multiplicative scanner effects remained after applying longitudinal and cross-sectional versions of ComBat. Cross-sectional ComBat and unharmonized data both showed residual scanner effects while longitudinal ComBat-adjusted data did not. Thus, we focused our attention on three cases: longitudinal ComBat with scanner fixed effects excluded (proposed method), and cross-sectional ComBat and unharmonized data, both with scanner fixed effects included. For each coefficient of interest, we compared the numbers of significant features, i.e., features with $p < 0.05/62$ (Bonferroni-corrected across features), across the three methods. Then, to avoid biasing our analyses to favor any particular method, we considered only features which were significant for all three cases. A good data harmonization method will ideally preserve the biological signal of interest while removing unwanted technical variability. Therefore, we might expect to see greater biological signal for the proposed method, in the form of greater magnitudes or smaller p-values for the longitudinal diagnosis-specific effects. Thus, for the statistically significant feature subsets, we compared coefficient magnitudes and p-values across the three methods.

4.2. Results

We found no significant additive or multiplicative scanner effects after applying longitudinal ComBat (Supplementary Tables S6 and S7). However, after cross-sectional ComBat, additive and multiplicative scanner effects were still significant for all features (Supplementary Tables S8 and S9).

For the AD $\times$ time coefficients, 30, 27, and 31 features passed the significance threshold for longitudinal ComBat, cross-sectional ComBat, and unharmonized data, respectively. Twenty-six features were significant for all three methods. Of these, 17 and 13 had larger coefficient magnitudes in longitudinal ComBat versus cross-sectional ComBat or unharmonized data, respectively, and 16 and 2 had smaller p-values in longitudinal ComBat versus cross-sectional ComBat or unharmonized data, respectively.

For the LMCI $\times$ time coefficients, 10, 11, and 11 features passed the significance threshold for longitudinal ComBat, cross-sectional ComBat, and unharmonized data, respectively. Nine features were significant for all three methods. Of these, 2 had larger magnitudes in longitudinal ComBat versus cross-sectional ComBat data; 3 had larger magnitudes in longitudinal ComBat versus unharmonized data; 3 had smaller p-values in longitudinal

versus cross-sectional ComBat data; and 1 had smaller p-values in longitudinal ComBat versus unharmonized data.

Coefficient estimates and corresponding KR p-values for the AD $\times$ time interaction are shown in Figure 3 and Supplementary Tables S10-S11. Coefficient estimates and corresponding KR p-values for the LMCI $\times$ time interaction are shown in Supplementary Figure S17 and Supplementary Tables S12-S13. Examples of unharmonized and harmonized trajectories are shown in Supplementary Figure S18.

5. Simulation Study

5.1. Methods

We performed a simulation study comparing the same approaches used above, longitudinal ComBat, cross-sectional ComBat, and unharmonized data, each with and without including scanner fixed effects in the model. For each iteration of the simulation, we began with the CN subset ($n = 197$) of the ADNI cortical thickness dataset. We randomly assigned each participant to either a CN control group or an AD group. In the AD group, for 6 of the 62 features, we added both an intercept and a slope effect to their cortical thickness trajectories. The magnitudes of these effects were estimated from the full ADNI dataset. We chose 2 strong, 2 moderate, and 2 weak effects (see Supplementary Table S14 for exact magnitudes of these effects). We then performed longitudinal ComBat and cross-sectional ComBat and fit the linear mixed effects model in Equation (1) (omitting the LMCI terms) to both harmonized and unharmonized datasets, with and without the scanner fixed effects in the model. The simulation was repeated for 1000 iterations.

We focused our primary analyses on estimation and inference for the AD $\times$ time coefficient. For the 56 null features, we compared distributions of the coefficient estimate means and standard errors over the 1000 simulations across methods. We also assessed distributions of type I error by calculating the percent of $p < 0.05$ from the KR test for each null feature. For the 6 features with nonzero effects, we compared distributions of the coefficient estimates and their standard errors. We assessed statistical power by calculating the proportion of $p < 0.05$ from the KR test. Finally, we looked at the distributions of intra-class correlation coefficients (ICC; $\widehat{\rho}/(\widehat{\rho} + \widehat{\sigma})$) for the nonzero features across each of the six methods. The ICC is a ratio of between-subject variation to total variation. Larger ICC is desirable because it allows for more clearly discernible between-subject differences.

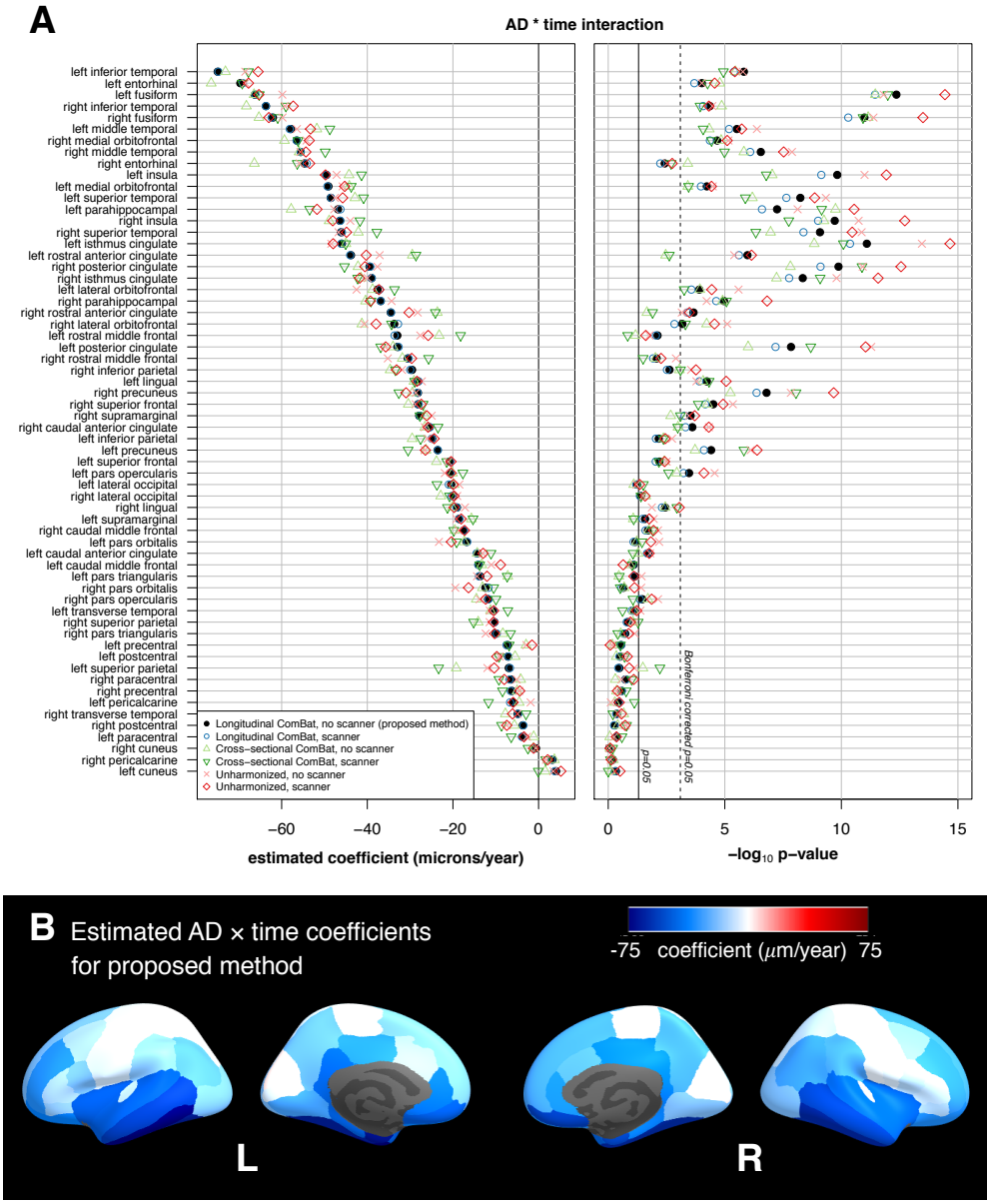

Figure 3: Comparison of data harmonization methods for the ADNI cortical thickness dataset. (A) Estimated coefficients and $-\log_{10} p$-values for the AD $\times$ time coefficients. Plots show results for each harmonization method, with and without scanner included as a fixed effect covariate in the final models. Features are sorted by coefficient magnitude for the proposed method — longitudinal ComBat with no scanner covariate in the final model. (B) Estimates obtained from data harmonized using the proposed method are displayed on the inflated cortical surface.

5.2. Results

For the 56 null features, Figure 4A shows that the distribution of the means of these coefficient estimates tend to be clustered more closely around zero for longitudinal ComBat as compared with the other methods, regardless of whether scanner was included as a covariate. Additionally, standard errors tended to be lower for longitudinal ComBat and unharmonized data methods than for cross-sectional ComBat. Longitudinal Combat resulted in type I error closer to the nominal rate than the other methods, ranging from 1.8 to 8.6% and 0.6 to 6.5% (below 5% for 22 and 49 of 56 features) when scanner was omitted or included in the final model, respectively. In contrast, type I error for unharmonized data was 4.3 to 14.5% and 4.5 to 13.1% (below 5% for 1 and 2 of 56 features) when scanner was omitted or included in the final model, respectively.

Results showed a similar pattern among the six nonzero features. Figure 4B shows results for one of each of the different effect sizes. (See Supplementary Figures S19 and S20 for results from the other nonzero features.) Again, the longitudinal ComBat methods resulted in the smallest standard errors in most cases, while cross-sectional ComBat tended to show the largest standard errors. Longitudinal ComBat was more powerful in detecting a weak effect size, correctly rejecting the null (uncorrected $p < 0.05$) in 80.1 and 75.0% of cases when scanner was omitted or included, respectively, versus 55.0 and 51.6% for cross-sectional ComBat, and 78.3% and 77.9% for unharmonized data. For all six features, the ICC was smaller after cross-sectional ComBat than after longitudinal ComBat or in unharmonized data, indicating relatively more within-subject and less between-subject variation (Supplementary Figure S21).

6. Discussion

In this study, we proposed and validated longitudinal ComBat, a method for harmonizing longitudinal data across different scanners. We first demonstrated methods for quantifying and visualizing additive and multiplicative scanner effects in unharmonized longitudinal data. We showed the presence of scanner effects in an ADNI-1 cortical thickness dataset for all of the 62 brain regions considered. Additive scanner effects were most prominent in medial occipital and medial temporal brain regions, while multiplicative scanner effects were strongest in superior frontal, rostral middle frontal, and superior parietal regions.

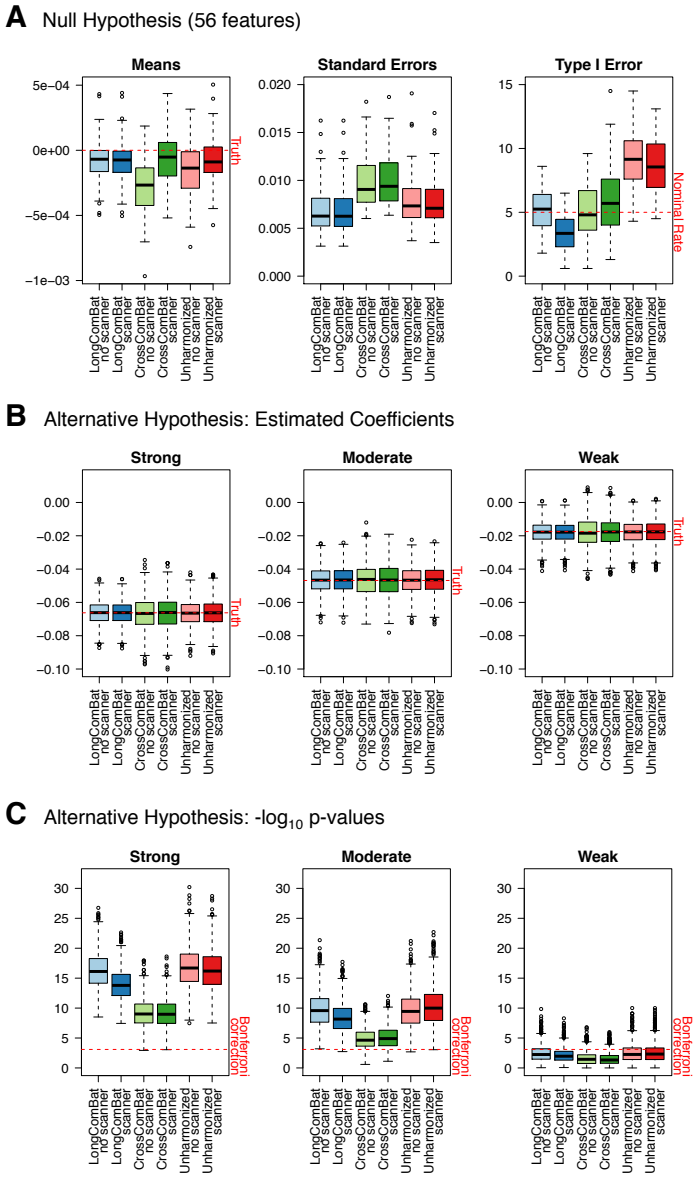

Figure 4: Simulation study results for three harmonization methods, each without or with scanner fixed effect covariates in the model. (A) Boxplots show distributions of the mean AD $\times$ time coefficient estimates over 1000 simulations for the 56 null features (left), the standard errors of the estimates (center), and the percentage of p-values < 0.05 from the Kenward-Roger test (right). (B) Distributions of the AD $\times$ time coefficient estimates over 1000 simulations for one strong, one moderate, and one weak effect size. (C) Distributions of the corresponding $-\log_{10}$ Kenward-Roger p-values.

Fourty-four of the 58 sites included in our dataset used more than one scanner, or upgraded scanners over the course of the study. We found that incorporating scanner information accounted for significantly more variability in the data than site alone. This is not surprising, as some sites used both 1.5 and 3.0 T scanners. It has been previously reported that higher field strengths tend to generate larger cortical thickness estimates (Han et al., 2006), which is consistent with our results. Even so, our results and other studies indicate that scanner effects have other sources beyond field strength (Han et al., 2006; Gunter et al., 2009; Lee et al., 2019). For example, Lee et al. (2019) analyzed ADNI participants scanned at 1.5 T, and quantified the effects of specific intra- and inter-vendor scanner changes on percent brain volume change measures. The authors found that inter-vendor and pulse sequence changes had the largest effects, as did, to a lesser extent, intra-vendor scanner upgrades. Thus, when seeking to minimize effects of scanner-induced variability in multi-scanner analyses, specific information about scanner hardware, acquisition parameters, and protocols should be taken into account whenever possible.

Moreover, we found that in this dataset, multiplicative scanner effects showed a relationship with scanner vendor. While much work went into standardizing protocols across sites and platforms in ADNI-1 (Jack Jr et al., 2008), technical variability was not completely eliminated. For example, Gunter et al. (2009) report that longitudinal analyses of the ADNI phantom revealed that, prior to mid-2007, autoshim was incorrectly disabled for one vendor protocol. This was later corrected. It is not immediately clear how this and other inter-vendor differences might have impacted the current dataset. In any case, the proposed harmonization method may be applied without explicit knowledge of the mechanisms underlying the mean shift or heteroscedasticity across scanners. However, a limitation of our methods is that our definition of scanner, as unique combinations of study site, scanner vendor, head coil, and field strength, may have missed hardware changes of the same model. Additionally, we were unable to account for changes in acquisition protocol such as the autoshim status, as Gunter et al. (2009) report that this was not recorded in DICOM headers. This highlights the importance of carefully tracking any hardware, software, or protocol changes in longitudinal imaging studies.

Notably, the cross-sectional version of ComBat we implemented, which does not account for within-subject repeated measures, did not completely remove additive and multiplicative scanner effects, and in fact tended to ex-

acerbate multiplicative scanner effects. Our proposed longitudinal version of ComBat, however, successfully removed both types of scanner effects. Perhaps it would be more fair to use only a cross-sectional subsample of the data to estimate scanner effects, as the cross-sectional method we applied violated the linear model assumption of independent residuals. However, as found by Venkatraman et al. (2015), when dependence is properly accounted for, there are advantages to using the entire longitudinal data to estimate scanner effects, as this allows one to decompose the within- and between-subject variability, and thus estimate scanner effects with greater precision. This may be particularly important when estimating and correcting for scanner-related heteroscedasticity.

We compared estimated rates of cortical atrophy in LMCI and AD participants, relative to controls, across three harmonization methods — unharmonized data, cross-sectional ComBat, and longitudinal ComBat. Coefficient estimates and p-values differed somewhat among the methods, but in general diagnosis-related increases in atrophy rate appeared in expected cortical regions across methods. For the AD by time interaction, the largest effect sizes occurred in bilateral inferior and middle temporal, entorhinal, fusiform, and orbitofrontal cortex. For the LMCI by time interaction, effect sizes were smaller but appeared strongest in similar regions, and particularly on the left side of the brain. This lateral asymmetry has been found in previous studies, although overall findings have been mixed (Shi et al., 2009; Derflinger et al., 2011; Meiberth et al., 2015). While it is somewhat difficult to unequivocally conclude that there was an increase in biological signal (e.g., in the form of greater coefficient magnitudes or smaller p-values) for longitudinal ComBat versus the other harmonization methods for this dataset, simulation study results suggest that longitudinal ComBat was likely to produce more accurate estimates and valid inference.

Our simulation study revealed that longitudinal ComBat produced estimates with smaller standard errors and greater power to detect true effects, particularly in the case of weak effect sizes, than cross-sectional ComBat. The unharmonized data tended to outperform cross-sectional ComBat on these measures as well. However, the unharmonized data showed inflated type I error rate under the null hypothesis, even when scanner was included as a fixed effect covariate in the final models. Meanwhile, longitudinal ComBat controlled the type I error at close to the nominal rate, and was particularly conservative when scanner was included in the final models.

We considered including a subject-specific random slope in our linear

mixed effects model, in addition to the subject-specific random intercept, but this accounted for relatively little variation in the data (only up to 0.7%). Thus, for the sake of parsimony, we chose to omit random slopes. However, it would be worthwhile to consider random slopes, along with more hierarchical versions of ComBat in the future. For example, it may be useful to incorporate information about site, field strength, or scanner vendor, in addition to scanner, so as to borrow information across scanners of similar type.

The proposed longitudinal version of ComBat would be useful for other types of longitudinal data requiring harmonization, such as genomic data, or neuroimaging studies of neurodevelopment, psychiatric disorders, or neurological diseases other than AD. One limitation of the method is that, in order to estimate scanner effects, it requires at least two scans per scanner. Despite this limitation, the method is flexible and may be applied to many existing and future longitudinal datasets. Code for implementing longitudinal ComBat is available at `https://github.com/jcbeer/longCombat`.

7. Acknowledgements

Funding: This work was supported by the National Institute of Neurological Disorders and Stroke grant numbers R01 NS085211 and R01 NS060910; the National Institute of Mental Health grant numbers T32 MH106442 and U01 MH109991; and a seed grant from the University of Pennsylvania Center for Biomedical Image Computing and Analytics (CBICA).

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

Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The authors thank Andrew Holbrook for helpful discussion.

Appendix A. Parameter estimation

Hyperparameters $\gamma_i, \tau_i^2, \lambda_i, \theta_i$ are estimated from standardized data using the method of moments. Let $\widehat{\gamma}_{i\nu} = \frac{1}{n_i}\sum_{jk} z_{ij\nu}(t_{ijk})$ (scanner i sample mean for feature v; note that these are on a different scale than the $\widehat{\gamma}_{i\nu}$ above), where $k \in \{1, \ldots, K\}$ is the visit index and n_i is total number of images from scanner i. Method of moments estimates for γ_i and τ_i^2 are

$$\overline{\gamma}_i = \frac{1}{V}\sum_{\nu} \widehat{\gamma}_{i\nu} \quad \text{and} \quad \overline{\tau}_i^2 = \frac{1}{V-1}\sum_{\nu} \left(\widehat{\gamma}_{i\nu} - \overline{\gamma}_i\right)^2.$$

Let $\widehat{\delta}_{i\nu}^2 = \frac{1}{n_i - 1}\sum_{jk} \left(z_{ij\nu}(t_{ijk}) - \overline{\gamma}_i\right)^2$ (scanner i sample variance for feature v). The sample mean and variance of the $\widehat{\delta}_{i\nu}^2$ can be calculated as

$$\overline{D}_i = \frac{1}{V}\sum_{\nu} \widehat{\delta}_{i\nu}^2 \quad \text{and} \quad \overline{S}_i^2 = \frac{1}{V-1}\sum_{\nu} \left(\widehat{\delta}_{i\nu}^2 - \overline{D}_i\right)^2,$$

respectively. We then set these sample moments equal to the theoretical moments of the Inverse Gamma distribution; the mean is $\theta_i/(\lambda_i - 1)$ and the variance is $\theta_i^2/\left[(\lambda_i - 1)^2(\lambda_i - 2)\right]$. Solving the system for λ_i and θ_i gives the estimates

$$\overline{\lambda}_i = \frac{\overline{D}_i^2 + 2\overline{S}_i^2}{\overline{S}_i^2} \quad \text{and} \quad \overline{\theta}_i = \frac{\overline{D}_i^3 + \overline{D}_i\overline{S}_i^2}{\overline{S}_i^2}.$$

Empirical Bayes estimates for scanner effect parameters $\gamma_{i\nu}$ and $\delta_{i\nu}^2$ are given by conditional posterior means. Let the conditional posterior distribu-

tion of $\gamma_{i\nu}$ be denoted by $\pi(\gamma_{i\nu}|\mathbf{Z}_{i\nu}, \delta^2_{i\nu})$. According to Bayes' Theorem,

$$\begin{aligned}
\pi(\gamma_{i\nu}|\mathbf{Z}_{i\nu}, \delta^2_{i\nu}) &\propto L(\mathbf{Z}_{i\nu}|\gamma_{i\nu}, \delta^2_{i\nu})\pi(\gamma_{i\nu}) \\
&\propto \exp\left\{-\frac{1}{2\delta^2_{i\nu}}\sum_{jk}(z_{ij\nu}(t_{ijk}) - \gamma_{i\nu})^2\right\}\exp\left\{-\frac{1}{2\tau_i^2}(\gamma_{i\nu} - \gamma_i)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{n_i\tau_i^2 + \delta^2_{i\nu}}{\delta^2_{i\nu}\tau_i^2}\right)\left[\gamma^2_{i\nu} - 2\left(\frac{\tau_i^2\sum_{jk} z_{ij\nu}(t_{ijk}) + \delta^2_{i\nu}\gamma_i}{n_i\tau_i^2 + \delta^2_{i\nu}}\right)\gamma_{i\nu}\right]\right\}.
\end{aligned}$$

By completing the square, we can identify the above as the kernel of a Normal distribution with expected value

$$E[\gamma_{i\nu}|\mathbf{Z}_{i\nu}, \delta^2_{i\nu}] = \frac{\tau_i^2\sum_{jk} z_{ij\nu}(t_{ijk}) + \delta^2_{i\nu}\gamma_i}{n_i\tau_i^2 + \delta^2_{i\nu}}.$$

This can be estimated using $\widehat{\gamma}_{i\nu}$, $\overline{\gamma}_i$, $\overline{\tau}_i^2$, as defined above, and $\widehat{\delta}^{2*}_{i\nu}$, defined below:

$$\widehat{\gamma}^*_{i\nu} = \widehat{E}[\gamma_{i\nu}|\mathbf{Z}_{i\nu}, \delta^2_{i\nu}] = \frac{n_i\overline{\tau}_i^2\widehat{\gamma}_{i\nu} + \widehat{\delta}^{2*}_{i\nu}\overline{\gamma}_i}{n_i\overline{\tau}_i^2 + \widehat{\delta}^{2*}_{i\nu}}.$$

Let the conditional posterior distribution of $\delta^2_{i\nu}$ be denoted by $\pi(\delta^2_{i\nu}|\mathbf{Z}_{i\nu}, \gamma_{i\nu})$. According to Bayes' Theorem,

$$\begin{aligned}
\pi(\delta^2_{i\nu}|\mathbf{Z}_{i\nu}, \gamma_{i\nu}) &\propto L(\mathbf{Z}_{i\nu}|\gamma_{i\nu}, \delta^2_{i\nu})\pi(\delta^2_{i\nu}) \\
&\propto \left(\delta^2_{i\nu}\right)^{\frac{n_i}{2}}\exp\left\{-\frac{1}{2\delta^2_{i\nu}}\sum_{jk}(z_{ij\nu}(t_{ijk}) - \gamma_{i\nu})^2\right\}\left(\delta^2_{i\nu}\right)^{-\lambda_i+1}\exp\left\{-\frac{\theta_i}{\delta^2_{i\nu}}\right\} \\
&= \left(\delta^2_{i\nu}\right)^{-\left(\frac{n_i}{2}+\lambda_i\right)-1}\exp\left\{-\frac{\theta_i + \frac{1}{2}\sum_{jk}(z_{ij\nu}(t_{ijk}) - \gamma_{i\nu})^2}{\delta^2_{i\nu}}\right\}.
\end{aligned}$$

This is an Inverse Gamma distribution with expected value

$$E[\delta^2_{i\nu}|\mathbf{Z}_{i\nu}, \gamma_{i\nu}] = \frac{\theta_i + \frac{1}{2}\sum_{jk}(z_{ij\nu}(t_{ijk}) - \gamma_{i\nu})^2}{\frac{n_i}{2} + \lambda_i - 1}.$$

This can be estimated using $\overline{\theta}_i$, $\overline{\lambda}_i$, and $\widehat{\gamma}^*_{i\nu}$ as defined above:

$$\widehat{\delta}^{2*}_{i\nu} = \frac{\overline{\theta}_i + \frac{1}{2}\sum_{jk}\left(z_{ij\nu}(t_{ijk}) - \widehat{\gamma}^*_{i\nu}\right)^2}{\frac{n_i}{2} + \overline{\lambda}_i - 1}.$$

Note that the estimates for $\widehat{\gamma}^*_{i\nu}$ and $\widehat{\delta}^{2*}_{i\nu}$ depend on each other. They can be estimated iteratively, for example by first substituting $\widehat{\delta}^2_{i\nu}$ for $\widehat{\delta}^{2*}_{i\nu}$ to obtain the estimate $\widehat{\gamma}^*_{i\nu}$, then plugging this into the formula for $\widehat{\delta}^{2*}_{i\nu}$, and so on. As Johnson et al. (2007) note in their supplementary material, this is a special case of the expectation-maximization algorithm and tends to converge rather quickly in less than 30 iterations.

References

Bakkour, A., Morris, J. C., Dickerson, B. C., 2009. The cortical signature of prodromal AD: regional thinning predicts mild AD dementia. Neurology 72 (12), 1048–1055.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67 (1), 1–48.

Conover, W. J., Johnson, M. E., Johnson, M. M., 1981. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23 (4), 351–361.

Derflinger, S., Sorg, C., Gaser, C., Myers, N., Arsic, M., Kurz, A., Zimmer, C., Wohlschläger, A., Mühlau, M., 2011. Grey-matter atrophy in Alzheimer's disease is asymmetric but not lateralized. Journal of Alzheimer's Disease 25 (2), 347.

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., et al., 2008. The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cerebral Cortex 19 (3), 497–510.

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

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., et al., 2017.

Harmonization of multi-site diffusion tensor imaging data. NeuroImage 161, 149–170.

Gunter, J. L., Bernstein, M. A., Borowski, B. J., Ward, C. P., Britson, P. J., Felmlee, J. P., Schuff, N., Weiner, M., Jack, C. R., 2009. Measurement of MRI scanner performance with the ADNI phantom. Medical Physics 36 (6, Part 1), 2193–2205.

Halekoh, U., Højsgaard, S., 2014. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. Journal of Statistical Software 59 (9), 1–30.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., et al., 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.

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27 (4), 685–691.

Johnson, W. E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8 (1), 118–127.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van Der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., et al., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. NeuroImage 30 (2), 436–443.

Kenward, M. G., Roger, J. H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 983–997.

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

Lee, H., Nakamura, K., Narayanan, S., Brown, R. A., Arnold, D. L., Initiative, A. D. N., et al., 2019. Estimating and accounting for the effect of MRI scanner changes on longitudinal whole-brain volume change measurements. NeuroImage 184, 555–565.

MATLAB, 2018. MATLAB version 9.4.0.813654 (R2018a). The Mathworks, Inc., Natick, Massachusetts.

Meiberth, D., Scheef, L., Wolfsgruber, S., Boecker, H., Block, W., Träber, F., Erk, S., Heneka, M. T., Jacobi, H., Spottke, A., et al., 2015. Cortical thinning in individuals with subjective memory impairment. Journal of Alzheimer's Disease 45 (1), 139–146.

Müller, C., Schillert, A., Röthemeier, C., Trégouët, D.-A., Proust, C., Binder, H., Pfeiffer, N., Beutel, M., Lackner, K. J., Schnabel, R. B., et al., 2016. Removing batch effects from longitudinal gene expression-quantile normalization plus ComBat as best approach for microarray transcriptome data. PloS ONE 11 (6), e0156594.

Murdoch Childrens Research Institute Developmental Imaging Group, 2017. freesurfer_statsurf_display: Freesurfer surface results display in MATLAB. `https://github.com/DevelopmentalImagingMCRI/freesurfer_statsurf_display`, accessed: 2019-11-25.

Orlhac, F., Boughdad, S., Philippe, C., Stalla-Bourdillon, H., Nioche, C., Champion, L., Soussan, M., Frouin, F., Frouin, V., Buvat, I., 2018. A postreconstruction harmonization method for multicenter radiomic studies in PET. Journal of Nuclear Medicine 59 (8), 1321–1328.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Shi, F., Liu, B., Zhou, Y., Yu, C., Jiang, T., 2009. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. Hippocampus 19 (11), 1055–1064.

Takao, H., Hayashi, N., Ohtomo, K., 2011. Effect of scanner in longitudinal studies of brain volume changes. Journal of Magnetic Resonance Imaging 34 (2), 438–444.

Tustison, N. J., Holbrook, A. J., Avants, B. B., Roberts, J. M., Cook, P. A., Reagh, Z. M., Duda, J. T., Stone, J. R., Gillen, D. L., Yassa, M. A., et al., 2019. Longitudinal mapping of cortical thickness measurements: An Alzheimer's Disease Neuroimaging Initiative-based evaluation study. Journal of Alzheimer's Disease 71 (1), 165–183.

Venkatraman, V. K., Gonzalez, C. E., Landman, B., Goh, J., Reiter, D. A., An, Y., Resnick, S. M., 2015. Region of interest correction factors improve reliability of diffusion imaging measures within and across scanners and field strengths. NeuroImage 119, 406–416.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., et al., 2015. 2014 update of the Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimer's & Dementia 11 (6), e1–e120.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., Sheline, Y. I., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Human Brain Mapping 39 (11), 4213–4227.