Site Harmonization Using Average Treatment Effect for the Overlap Population

Jongmin Mun
Data Sciences and Operations Department
Marshall School of Business
University of Sothern California
Los Angeles, CA 90089
jongminm@marshall.usc.edu

Jaeoh Kim
Department of Data Science
Inha University
Incheon, South Korea
jaeoh.k@inha.ac.kr

Bo-yong Park
Department of Brain and Cognitive Engineering
Korea University
Seoul, South Korea
boyongpark@korea.ac.kr

Abstract

Pooling multi-site MRI data is crucial for enhancing statistical power in neuroimaging studies. However, site-related variability introduces significant challenges, particularly when biological covariates are unevenly distributed across sites. Traditional harmonization methods, such as ComBat and its extensions, rely on strong modeling assumptions about the conditional expectation of imaging features given site and covariates, which may not scale effectively to high-dimensional settings. In this study, we reframe site harmonization as a causal inference problem and propose the use of overlap weighting, an improved variant of inverse propensity score weighting (IPW), to estimate and remove site effects. Unlike conditional outcome modeling, overlap weighting does not require direct modeling of high-dimensional outcomes and naturally accommodates heterogeneous site effects. It also mitigates the issue of extreme weights that often hampers IPW. Applying our method to the ABIDE1 dataset, which includes MRI features from 748 brain regions across 19 sites, we show that ATO-based harmonization enhances detection power in group comparisons relative to ComBat, while preserving valid statistical inference.

1 Introduction

Most MRI studies have relatively small sample sizes. To overcome limitations in statistical power and improve generalizability, neuroimaging researchers are increasingly pooling MRI data from multiple sites. However, aggregating data across sites introduces unwanted variability, particularly due to differences in scanner hardware and acquisition protocols, which must be mitigated through post-acquisition harmonization. A major challenge arises because biological covariates such as age and sex often differ across sites, and these covariates can significantly influence MRI measurements. Thus, effective harmonization must eliminate site effects while accounting for differences in covariate distributions.

Preprint. Under review.

2 Previous Literature

One of the most widely used harmonization methods in the neuroimaging community is ComBat (Johnson et al., 2007), a mixed-effects model that accounts for both additive and multiplicative site effects. Formally, the ComBat model is given by:

$$y_{ijv} = \alpha_v + \mathbf{X}_{ij}^\top \boldsymbol{\beta}_v + \gamma_{iv} + \delta_{iv}\varepsilon_{ijv}, \tag{1}$$

where i indexes the site, j the participant, and v the brain region of interest (ROI). In this model, x_{ijv} denotes the measurement of the v-th brain ROI for participant j from site i; γ_{iv} and δ_{iv} capture additive and multiplicative site effects, respectively. The residual term ε_{ijv} is assumed to be normally distributed with mean zero and variance σ_v^2. The covariate vector $\mathbf{X}_{ij}$ typically includes variables such as age and sex. In addition, it assumes that the additive and multiplicative effects of the sites are not completely independent across ROIs but, rather, they share a common distribution. Such considerations prevent the use of standard linear models, but ComBat uses an empirical Bayes framework to estimate the distribution of the effects of site.

Efter estimation, the adjusted measurement using ComBat is then:

$$y_{ijv}^{\text{ComBat}} = \frac{x_{ijv} - \hat{\alpha}_v - \mathbf{X}_{ij}^\top \hat{\boldsymbol{\beta}}_v - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + \hat{\alpha}_v + \mathbf{X}_{ij}^\top \hat{\boldsymbol{\beta}}_v. \tag{2}$$

Due to its linear model structure, ComBat and its extensions are limited in capturing complex, nonlinear site differences or covariate effects. To address this limitation, machine learning-based methods have recently emerged. For example, An et al. (2025) propose DeepResBat, which generalizes the ComBat model as follows:

$$y_{ijv} = f_v(\mathbf{X}_{ij}) + g_v(i), \tag{3}$$

where f_v models nonlinear covariate effects using tree-based methods, for example, XGBoost (Chen and Guestrin, 2016), and $g_v(i)$ represents the residual site effect estimated via a conditional variational autoencoder.

3 Causal Inference Approach

Causal inference approaches to site harmonization remain relatively scarce. However, both ComBat and DeepResBat can be interpreted through the lens of Conditional Outcome Modeling (COM) in causal inference. In this framework, these methods aim to directly estimate the conditional expectation $\mathbb{E}[Y \mid T, X]$, where T represents the site (analogous to the treatment assignment) and X denotes the covariates. A common assumption in causal inference is that each subject j has a set of potential outcomes, one for each possible site (or treatment level in causal terminology):

$$Y_{jv}(1), Y_{jv}(2), Y_{jv}(3), \ldots,$$

where the number in the parentheses indicates the site. Under standard assumptions, such as conditional unconfoundedness and positivity, we can identify the causal effect of site, a quantity known as the Average Treatment Effect (ATE), by aggregating conditional expectations. For instance, the ATE between site 1 and site 2 is given by:

$$\mathbb{E}_X \left[\mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=2, X] \right].$$

In this framework, ComBat assumes a linear form for $\mathbb{E}[Y|T, X]$, while DeepResBat assumes a nonlinear relationship. It is worth emphasizing that both methods implicitly assume the treatment effect is the same across individuals, which may not be realistic. In addition, when the measurements are high-dimensional, as in neuroimaging data, fitting such models becomes computationally expensive.

An alternative approach from the causal inference literature is Inverse Propensity Weighting (IPW), which creates a reweighted pseudo-population in which associations reflect causal effects. ATE estimation via IPW has several well-known conceptual and computational limitations, which may have hindered neuroimaging researchers from adopting this method for site harmonization. To address these issues, we focus on a more advanced statistical methodology: estimation of the Average Treatment Effect for the Overlap population (ATO) via overlap weights (Li, 2019). ATE estimation via IPW can be understood as a special case of this method.

To explain, first define the propensity score for site i as:

$$e_i(x) = \Pr(T_j = i \mid X_j = x),$$

which must be estimated from the data, typically via logistic regression. However, any modeling approach can be used, including more flexible methods such as XGBoost.

The expectation of the potential outcome $\mathbb{E}_X[\mathbb{E}[Y|T = 1, X]]$, in a reweighted pseudo-population under site i, can be estimated by:

$$\hat{m}_{iv} = \frac{\sum_{j=1}^{n} Y_{ijv} w_i(X_{ijv})}{\sum_{j=1}^{n} w_i(X_{ijv})}, \tag{2.6}$$

where the overlap weight is defined as:

$$w_i(X) = \left(\sum_{k=1}^{J} \frac{1}{e_k(X)} \right)^{-1} \frac{1}{e_i(X)}, \quad i = 1, \ldots, J,$$

where J is the total number of sites. For example, the site effect between site 1 and site 2 can be estimated by $\hat{m}_1 - \hat{m}_2$. In contrast, standard IPW uses the weight $1/e_i(x)$, and the resulting $\hat{m}_1 - \hat{m}_2$ corresponds to the ATE.

For harmonization, we can simply compute

$$y_{ijv}^{\text{ATO}} = y_{ijv} - \hat{m}_{iv}. \tag{4}$$

Conceptually, ATO places more weight on observations with balanced covariates, that is, those with propensity scores closer to uniform, which resemble samples from a randomized controlled trial. This leads to more plausible causal estimates. In practice, ATO also reduces variance compared to IPW, which is often unstable in observational data due to extreme weights, as propensity scores are frequently estimated to be close to 0 or 1.

Li et al. (2018) show that under certain technical conditions in the two-site setting, ATO using overlap weights achieves consistency and efficiency advantages over ATE estimated via IPW.

The key advantage of ATO over COM-based methods such as ComBat and DeepResBat is that it avoids modeling the outcome directly, making it especially suitable for high-dimensional outcomes such as neuroimaging data. Moreover, while COM methods assume homogeneous treatment effects, ATO allows for heterogeneous site effects across individuals and aggregates these effects over the overlap population.

4 Data analysis

We used the ABIDE1 dataset, which includes MRI data from individuals with autism and healthy controls collected across 19 sites ($N = 621$). A total of 374 brain regions (ROIs) were defined, and one feature was extracted from each region for both structural and functional modalities, yielding $P = 748$ features per subject. The covariates used in our analysis are `AGE_AT_SCAN` and `SEX`. Table 1 summarizes the sample size for each site.

Table 1: Number of subjects per site in the ABIDE1 dataset (part 1)

LEUVEN_1	KKI	UM_1	OHSU	TRINITY	CMU	SDSU	PITT	USM	MAX_MUN
16	38	32	13	44	18	13	43	65	39

(continued)

NYU	UCLA_2	UM_2	OLIN	YALE	LEUVEN_2	UCLA_1	CALTECH	SBL
143	13	17	17	27	17	38	23	5

Figure 1-(a) compares the adjusted p-values of `DX_GROUP` obtained from standard linear model, after ComBat and ATO harmonization, both corrected using the Benjamini-Hochberg procedure. The plot shows that, for most hypotheses, ATO yields smaller p-values than ComBat. Figure 1-(b) illustrates that all hypotheses rejected by ComBat are also rejected by ATO, with ATO identifying slightly more rejections overall—366 compared to 360 for ComBat. Figure 2 presents the adjusted p-values from both methods for each individual hypothesis.

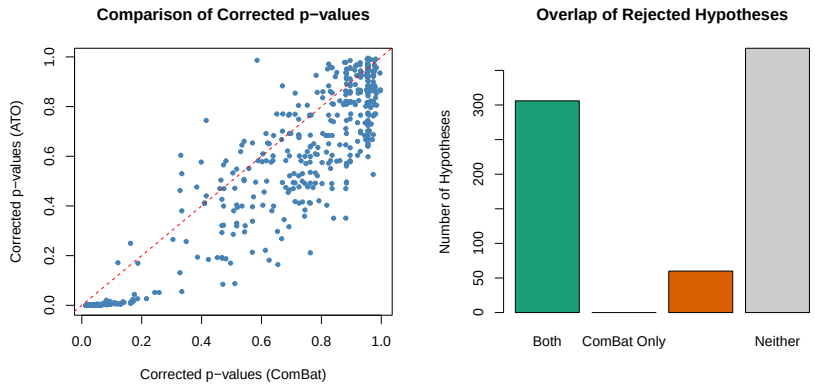

Figure 1

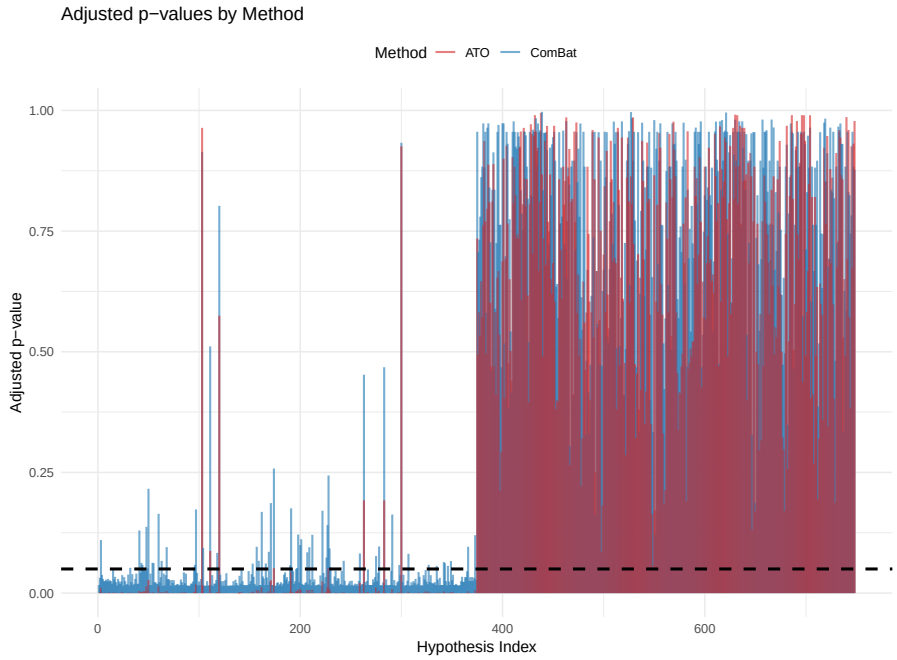

Figure 2

References

An, L., Zhang, C., Wulan, N., Zhang, S., Chen, P., Ji, F., Ng, K. K., Chen, C., Zhou, J. H., and Yeo, B. T. T. (2025). DeepResBat: Deep residual batch harmonization accounting for covariate distribution differences. *Medical Image Analysis*, 99:103354.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

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

Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389–2415.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400.